# Basic Survey Design

- Introduction
- The Set up Stage of a Survey
- Samples and Censuses
- Data Collection Methods
- Frames & Population
- Errors in Statistical Data
- Sample Design
- Questionnaire Design
- Survey Testing
- Design Data Processing
- Analysis
- Presentation of Results
- Confidentiality

## INTRODUCTION

The following manual is intended to provide an understanding of the issues involved in survey design. It provides the key issues to be considered when designing surveys and potential survey designs and covers the advantages and disadvantages of the various methods available.

By reading these guidelines you would not be expected to be a "sample design expert". You should however be able to:

- formulate survey objectives to discuss with consultants,
- identify appropriate survey techniques to collect data,
- design and test simple questionnaires.

The information provided in this document is by no means exhaustive, and you are encouraged to refer to relevant books and journals to follow up areas of particular interest.

## DEFINITIONS OF CONCEPTS USED WHEN COLLECTING DATA

The aim of this section is to enable you to become familiar with some of the more basic concepts involved in survey design. We will start off by providing a few basic definitions.

**Survey**
Surveys are used to collect information in order to answer a question or to make a decision. More formally, they are used to put a value onto some indicator or measure. Surveys measure one or more characteristics of a population. These characteristics may be measured by surveying all members of the population, or a sample of the population. A sample survey is a survey of a subset of the population.

**Census**
A census is an enumeration of the whole population. (See Samples and Censuses for more details.)

**Sample**
The sample is the set of observations which is taken from the population for the purpose of obtaining information about the population. (See Samples and Censuses for more details.)

**Unit**
A unit is the base level at which information is sought. There are two types of units: selection units and reporting units. A selection unit is a unit that is selected in the survey and a reporting unit is a unit that reports the information required. The reporting unit may not be the same as the selection unit (eg. in a household survey, the selection unit may be a household, but the reporting unit may be any responsible adult living in that household).

**Population**
The population is the complete set of units from which information is to be obtained and about which inferences are to be made. There are two types of population: target population and survey population. A target population is the population about which information is to be sought and a survey population is the population from which information can be obtained in the survey. The target population is also known as the scope of the survey and the survey population is also known as the coverage of the survey. (See The Set-up Stage of a Survey for more details.)

**Frame/Sampling Frame**
The frame is the list of all selection units in the survey population. (See Frames & Population for more details.)

**Population Parameter**
A population parameter is a summary measure of a population, the value of which helps to describe a population. An example of this is average weekly income, which is often used as an indicator of well-being or of spending power in the community.

**Sample Statistic**
A sample statistic is a summary value of a variable calculated from the sample.

## DEFINITIONS OF CONCEPTS USED TO PRESENT DATA

From a sample of selected observations, we calculate certain statistics. Based on these statistics, we then make inferences about the population. For example, wine tasting is a form of sampling. A small sample of wine is taken, and from that sample, inferences are made about the contents of the whole vintage.

**Mean**
This refers to the average of a set of sample or population values. For example, the mean of 6, 7, 12, 15 and 20 is; (6 + 7 + 12 + 15 + 20)/5. The 5 corresponds to the number of values added together.
The mean therefore equals 12.

**Variability**
The spread of the values of a variable.

**Variance**
A measure of variability which measures the spread of the values around the mean. A larger variance indicates the data is more spread out around the mean.

**Standard Deviation**
Standard Deviation refers to the population variability based on the square root of the population variance.

**Standard Error**
Standard Error refers to establishing confidence levels in the data, based on the variability of a sample statistic.
The mean, variability and related topics are discussed in more detail in Analysis.

This page last updated 1 March 2023

# THE SET-UP STAGE OF A SURVEY

## Survey Set-up

The set-up stage is a very important part of any research, yet its value is often underestimated. Most research projects (and therefore, most surveys) which go badly wrong, do so because of inadequate attention given to setting the project up properly. For this reason, any effort to collect information, should not occur until the issues discussed in this Section are dealt with.

This Section is intended to assist researchers and managers in identifying the key factors in planning the research project, analyse their requirements and select the most appropriate method of collecting the necessary information.

## Aims and Objectives

The first and most important step, in the collection of any information, is to establish the research objectives clearly. From this basis, it is then possible to translate these objectives into establishing specific information requirements.
The objectives should define what the survey results will be used for and may include such uses as the basis for decision-making, the allocation of funds, to analyse the outcome of policies or programs (program evaluation) or to determine the direction of future operations. It is highly recommended that the number of objectives be kept to a minimum. With too many, the information gathering exercise can easily become too unwieldy and costly.

## The Target Population

Once our uses for the data have been determined, our next requirement is to define the target population. The target population is the group about which we wish to make inferences from the survey data. The target population should be defined in terms of :

- content (eg. all persons),
- units (eg. in households),
- extent (eg. in Australia), and
- time frame (eg. 6 September 1991).

An examination of the examples listed in brackets above reveals the target population of the 1991 Census of Population and Housing. A clear definition, such as that shown above, will give the researcher some idea of the study's potential size, as well as some insight into possible survey methodologies.

Note that the target population may not be easily accessible due to logistical or financial restraints. For example, in the Census of Population and Housing, it is difficult to obtain data on the homeless or people who are in transit from one place to another during Census day.

When considering the population that is to be studied, an important factor is the units that are to respond to the collection (*the reporting units*) and the units that are to be selected to access the reporting units (*the selection units*). An example of the difference between reporting units and selection units is given in Aims and Objectives, where the selection unit is a household and the reporting unit is any responsible adult resident at that household.

Methods of accessing populations are dealt with further in Frames & Population.

## Conducting Preliminary Research

This step involves becoming familiar with the issues to be researched. By doing so, researchers obtain a stronger focus for their study, as well as some insights on how best to collect the desired information. Preliminary research is done by gathering together relevant reports, information papers, academic journals, books, newspapers and government files. Similarly, talking to people who are familiar with the issues should also be undertaken.

## Establishing Data Requirements

The first consideration when clarifying information requirements is to clearly define the data items to be collected. Only with clear definitions of the terms and concepts involved can the researcher be assured that the data will be relevant to the project's aims and objectives. Standard definitions for data items should be investigated as the use of standards allows comparability between the results from the current investigation and previous data.

A common way to ensure all the appropriate data items are collected is to specify, as early as possible in the planning stages of the collection, the detailed tables to be derived. Such specifications not only help clarify the data items to be collected, but the levels or subgroups for which the data is required. Assuming an appropriate collection methodology, defining the levels of interest beforehand will help to ensure an adequate sample size is chosen to achieve the desired levels of accuracy.

The definition of concepts also plays a part in identifying the population. The survey objectives define the target population, but the specific data items and output requirements determine the make-up of subgroups inside the population. These subgroups make up the cells on the tables that have been determined in describing the output requirements.

All data items proposed for the survey should be scrutinised closely to establish whether they are consistent with the aims and objectives of the project. It is vital that extra information is not sought because it is of interest or is related to the issues being examined. This may lead to the survey losing its focus and add considerably to the cost and time to do the project.

Ideally the information requirements, whether they be factual data (eg. unemployment or inflation rates) or attitudinal data (eg. opinions on a particular topic), and the proposed methods of analysis should be put down in writing, as this can help to focus the researcher's attention on important considerations affecting the design of the collection.

### Financial and Time Constraints and Other Considerations

Apart from the important considerations in survey design outlined above, there are several other aspects that need to be accounted for. The financial and time constraints placed on a survey are often a deciding factor influencing survey design, as well as the size of the survey and the level of accuracy sought. It is therefore important to make both time and financial estimates for each aspect of a survey and monitor them closely.

Other aspects include creating or maintaining the survey frame, designing the sample selection, choosing a data collection method, designing the questionnaire and determining the field and data processing procedures.

## NON-SURVEY METHODOLOGIES

As described above, the characteristics of our data items often point at the best way to collect our information. How expensive and time-consuming the data are to collect are also often important considerations when choosing whether to run a survey. The availability of similar data from other sources also plays a part, but collating data from many sources may also prove time-consuming and expensive.

### Existing Data Sources

The information that is required may already be available from another data source. Research of subject matter documentation, as discussed in Conducting Preliminary Research above, may provide the necessary information on some topics. Some data may also be available from government departments and agencies as a by-product of their administrative information (eg. Registrar of Births, Deaths and Marriages). Previous surveys may also have been carried out in the field of interest. Care must be taken that the recording of the data comes from reliable sources and that the definitions of the data items used closely approximate those currently of interest. This is a cheap and fast way of obtaining data so long as the existing data meet the requirements of the research.

### Case Study

In preparing a case study the researcher seeks to collect and analyse as much data about the chosen subject as possible from a relatively small number of cases. Case studies allow greater detail to be gathered for less cost than in most surveys, but the data is less likely to reflect the whole population.

### Participatory Observation

Participant observation is a method of data collection whereby the researcher seeks to become a member of the group under study, and records data from within. This method is often used in qualitative studies ie. studies that deal with variables which cannot be readily measured such as peoples perception, group dynamics, etc.

### Advantages and Disadvantages

The main advantages of non-survey methodologies is that they provide data, often quite detailed, in-depth data, at a relatively low cost and that they can result in substantial reductions in the time taken to produce results. However, statistical accuracy may be traded off to achieve these gains. Existing data sources may use definitions that do not match the required definitions very well, and the other non-survey methodologies use only very small numbers of subjects in the study and are open to the subjects not being representative of the target population.

This does not mean that non-survey methodologies should be overlooked entirely - it only means that care should be taken when analysing their results. For example, if we chose to look at the effect of overcrowding in our hospitals, we may decide to perform a case study or control group experiment rather than a survey as these methods allow the researcher to collect information for a comparatively modest cost.

## SURVEY DESIGN CONSIDERATIONS

After considering all the options available to you, should you decide that a survey is the option most suited to your study, your next task will be to decide upon the survey design.

The three objectives of sample surveys are to

- use the data collected from a sample of population units (using the sample statistics calculated),
- to describe a population, i.e. estimate certain population parameters (descriptive statistics), and
- to test statistical hypotheses about the population.

Issues relating to the survey design encompass all the methodological and organisational aspects of the survey to be conducted and involve making decisions about how these will be carried out. In doing this, decisions that are technically desirable need to be balanced off against what is practically feasible, and need to take into account the purposes of the survey, the accuracy required in the results, the resources available (ie financial, timing, staffing, and other resources) and other practical considerations. The aim of survey design is to best meet the aims of the survey within the boundaries of the resources that are available. Survey design issues to be considered include :

- the survey objectives,
- the population of interest (or target population),
- the reference period for the data,
- Geographic, demographic boundaries,
- the frame and the units,
- the sample design and sample size,
- the data collection method,
- the questionnaire,
- non-sampling errors,
- non-response,
- the field procedures and data processing system, and
- the tabulations and analyses required.

Some of these points have been dealt with in this Chapter. The rest are dealt with in the remaining Chapters.

## SURVEY MANAGEMENT

Survey management involves organising and controlling each aspect of the survey. The staff responsible for the survey need to be skilled and experienced in order to avoid the many pitfalls that confront the researcher new to the field. The operational aspects of a survey require thorough planning and efficient management of financial and human resources. In the course of addressing survey management issues, a researcher may find that the survey needs to be modified before it can be conducted feasibly. Since survey management is an integral part of a survey, sufficient resources should be allocated to it to ensure that the survey runs smoothly and produces reliable and timely results.

### Advertising

There are several parts of the survey where advertising is necessary. These include advertising for staff which can be done through the daily newspapers or through Centrelink and pre-survey publicity to potential respondents which can be done through an introductory letter sent to respondents. After the survey, the availability of survey results can be advertised through the media or by contacting interested parties. Publicity of the implementation of survey results can also raise the usefulness of the study.

When advertising for staff, you should specify what the job is, subject of the survey, amount of interviewing work, need for prior experience if required, rate of pay, number of days interviewer is required for, whether evening/weekend work is involved and area where interviews will be conducted.

### Finance

The financial constraints placed on a survey is often a deciding factor influencing the survey. It is therefore important to make financial estimates for each aspect of a survey and to monitor spending (and therefore the progress of the survey) closely.

Adequate resources should be allocated to survey management as it plays a central role in the conduct of a survey, particularly the most labour intensive part. Financial estimates for other aspects of a survey generally fall into three categories:

- overheads,
- salaries, and
- survey and processing costs.

Some areas requiring financial estimates includes advertising, hire of rooms, power, office equipment and supplies, printing of forms and questionnaires, postage, salaries of office and field staff, survey testing, travel expenses, data processing and publication and dissemination. All of these estimates need to be made at an early stage in the development of the study and also monitored along the way.

### Recruitment

Regardless of particular collection methods used, personnel need to be recruited and trained to collect, despatch and process the survey questionnaires as well as interviewing respondents. When recruiting staff, it is necessary that they have the basic skills and attributes required to do the work, and in the case of a national survey, they must be in or near the areas where they are expected to work. The number of staff that you need to recruit depends on the size of the sample, the length and complexity of questionnaire, the timetable and the availability of funds. Staffing levels may vary during the research period depending on staff retention rates or whether there are peak periods or low periods.

### Training

The adequacy of training given to interviewers and processing staff has a strong influence on the quality of results obtained from the survey. Thorough training of interviewers is important because they have a wide range of tasks to perform and are the main point of contact between the respondents and the researcher. Comprehensive training of office staff should enable them to process the survey questionnaires as accurately and as quickly as possible. Training can be provided in the form of:

- manuals,
- formal training courses, and
- "on the job" training.

It is also useful to include some form of home study exercise to test understanding of instructions and survey procedures. Interviewers and processing staff will improve as they gain practical experience "on the job" and consolidate their formal training.

### Interviewer Training

Training for interviewers includes the purpose of the survey; the scope and coverage of the survey; a general outline of the adopted sampling approach; the format of the questionnaire; recording of the responses; correct interviewing techniques; techniques for avoiding or reducing non-response and maintaining respondent co-operation; field practice; quality control; editing; planning the workload; and administrative arrangements.

### Training for Processing Staff

For processing staff the training should cover the purpose of the survey; the scope and coverage of the survey; the questionnaire; the way the responses are recorded; coding; editing; key-punching instructions if the survey responses are to be punched onto a computer database; quality control; and administrative arrangements.

This page last updated 1 March 2023

## SAMPLES AND CENSUSES

Surveys are used as a tool to collect information from some or all units of a population and compile the information into a useful form. There are two different types of surveys that can be used to collect information in different circumstances to satisfy differing needs. These are sample surveys and censuses.

### SAMPLE SURVEYS

In a sample survey, only part of the total population is approached for information on the topic under study. These data are then 'expanded' or 'weighted' to make inferences about the whole population. We define the sample as the set of observations taken from the population for the purpose of obtaining information about the population.

Advantages of Sample Surveys compared with Censuses:

- Reduces cost - both in monetary terms and staffing requirements.
- Reduces time needed to collect and process the data and produce results as it requires a smaller scale of operation.
- (Because of the above reasons) enables more detailed questions to be asked.
- Enables characteristics to be tested which could not otherwise be assessed. An example is life span of light bulbs, strength of spring, etc. To test all light bulbs of a particular brand is not possible as the test needs to destroy the product so only a sample of bulbs can be tested.
- Importantly, surveys lead to less respondent burden, as fewer people are needed to provide the required data.
- Results can be made available quickly

Disadvantages of Sample Surveys compared with Censuses:

- Data on sub-populations (such as a particular ethnic group) may be too unreliable to be useful.
- Data for small geographical areas also may be too unreliable to be useful.
- (Because of the above reasons) detailed cross-tabulations may not be practical.
- Estimates are subject to sampling error which arises as the estimates are calculated from a part (sample) of the population.
- May have difficulty communicating the precision (accuracy) of the estimates to users.

### Sample Survey Design Considerations

When running a sample survey, there are several design considerations that need to be taken into account that are specific to sample surveys. These additional factors include: sample size, sample design, the mode of estimation based on survey results and, where applicable, stratification, allocation of the sample across the strata and the selection of the sample within the strata. These factors, however, depend on many other factors such as the objectives of the survey, nature of target population, data items to be collected, level of accuracy required etc.

### CENSUSES

A census is a collection of information from all units in the population or a 'complete enumeration' of the population. We use a census when we want accurate information for many subdivisions of the population. Such a survey usually requires a very large sample size and often a census offers the best solution.

Advantages of Censuses compared with Sample Surveys:
The advantages of a census are that:

- Data for small areas may be available, assumimg satisfactory response rates are achieved.
- Data for sub-populations may be available, assumimg satisfactory response rates are achieved.
- (Because of the above reasons) detailed cross-tabulations may be possible.
- The estimates are not subject to sampling error.

An examination of these shows that these closely reflect the disadvantages of a sample survey.

Similarly, the disadvantages of censuses relate to the advantages of sample surveys.

This page last updated 1 March 2023

## DATA COLLECTION METHODS

There are a number of methods available to the researcher for collecting data. The commonly used data collection methods can be divided into three basic types:

- personal interview,
- self enumeration, and
- telephone interview.

These three manual, paper- based types have corresponding electronic methods:

- computer assisted personal interviewing (CAPI),
- computer assisted self- interviewing (CASI) which includes web and email forms, and
- computer assisted telephone interviewing (CATI).

A further data collection method not covered here is the use of existing data sources, such as administrative files. The success of the survey will depend to a large extent on the suitability of the data collection method chosen. The factors discussed below need to be considered when determining a data collection method that is appropriate for the survey requirements.

## FACTORS AFFECTING CHOICE OF COLLECTION METHOD

### Complexity of Topic

The depth and complexity of the topic to be covered and, in particular, the nature and form of the questions, will often dictate the collection method to be employed. The likely quality or length of the response sought may also influence the choice of collection method. For instance, it is often difficult to obtain answers to complex questions using telephone surveys, whereas self-administered surveys can include detailed explanatory notes, and personal interviews with the aid of such features as show-cards can often yield a more detailed response. For surveys with complicated sequencing, an electronic version that performs this sequencing well is preferable to a paper version.

### Types of data

The kind of topics that can be properly covered vary across data collection method. For sensitive topics, self-administered methods are generally preferred to interviews as they are seen as more private. Sometimes a face to face interview will include a self-administered module for this reason. Surveys which require respondents to check records before they can answer, or check with other people who may not be immediately available, should also use self-administered methods. For this reason a telephone survey might involve mailing a form out for respondents to complete, but ask them to read out their answers over the phone.

### Non-response

The quality and reliability of survey data can be significantly affected by the degree of response to a survey. Although it is very rare to achieve a 100 per cent response rate for any survey, the choice of collection method can significantly influence the response rate obtained. Personal interviews can achieve a better response rate than mail surveys or telephone surveys because it is more difficult to refuse to a person's face. However other factors such as whether the survey is compulsory, whether it is repeated and how the follow-up is conducted can mediate the effect of method on response.

### Respondent Preference

How the collection method fits in with the life style or working style of the respondent should influence what collection method will be used, as this will increase respondent motivation. For example, the Retail Business Survey works well as a CATI survey in part because the respondents do much of their business over the phone and therefore are comfortable with this method. Some respondents have a strong preference for completing their forms electronically, and since the Electronic Transaction Act (July 2001) came into effect, respondents can insist on their right to do this for ABS surveys.

### Resources Available

When human resources, money or time are limited, the researcher may choose to use a less appropriate collection method in order to boost the number of responses or get the responses more quickly. Such an option is often in conflict with the quality requirements of the survey. To compromise, a planned reduction in data quality, a reduction in coverage or using a less systematic way of choosing a sample, might be alternatives presented to the researcher. The researcher might also seek resources or cost savings elsewhere.

### Sampling Frame and the Population

The type and quality of the sampling frame may influence the choice of collection method. For example, if the target population includes the homeless, a mail survey is not suitable. The characteristics of the population may also influence the collection method. If we are interested in doing a survey which includes shift workers and other people who are not home during the times an interviewer could reach them, then a mail survey would be more appropriate than an interview.

## PERSONAL INTERVIEWS

### Personal Interview - Face to Face

In the ABS personal interviews are used mostly for household social surveys. Personal interviews involve a trained interviewer going to the potential respondent, asking the questions and recording the responses. Because of the interaction involved when the interviewer is present, this method of data collection is effective in terms of:

- establishing rapport,
- high response rates,
- conducting surveys which require complex sequencing.

Personal interviews are however very expensive to conduct. There are costs involved in personnel, time, and other resources required to obtain, train, and manage an interviewer workforce. Travelling from one respondent to another is another cost component in personal interview and is especially problematic when repeat visits are necessary to catch the respondent while they are available. Data can possibly be subject to bias caused by the interviewer's appearance and attitude as well the way they phrase and rephrase questions and explanations. Also, the respondents may not feel free to disclose sensitive or private information to an interviewer.

### Computer-Assisted Personal Interview

When performing a computer-assisted personal interview (CAPI), the interviewer takes a laptop computer to the interview and codes the data into the computer as it is provided. This technique allows;

- more flexibility to move around the form and skip questions
- some edit checks to be carried out at the time of the interview, thus improving data quality, and
- (because of the above reasons) the overall timeliness of the survey is improved.

Respondents can also sometimes take the survey more seriously because a computer is involved. However, the method does have several drawbacks: there is an increased cost in set-up, maintenance of the computer equipment and training of interviewers in the use of CAPI, and the actual interview time increases as responses are coded and edited at the time of the interview.

## SELF-ENUMERATION METHODS

In the ABS self-enumeration methods are mostly used in surveys of businesses. Self-enumeration surveys are those in which it is left to the respondents to complete the survey questionnaires. The following are three of the most common self-enumeration methods.

### Postal/Mail Survey

Questionnaires are mailed out to respondents with a return-paid envelope so that the respondent can mail back the completed form. This technique allows:

- the respondent is able to complete the questionnaire in their own time and not all in one session,
- the respondent can then check records and with other people,
- more detailed instructions and explanations can be included for better quality data,
- wider geographic areas can be covered for far lower cost than personal interviews,
- allows access to 'difficult-to-contact' respondents (those people who only provide post office boxes as addresses or are away from home a great deal).

The major disadvantages with postal surveys is that they can have a lower response rate, due in part to the difficulty in maintaining accurate address and contact details. The reliance on postal services and respondents completing questionnaires can cause data to be slow to become available due to the lag in time between when the questionnaire is mailed out and the time the questionnaire is returned, although this effect is reduced when surveys are repeated regularly. There is also a limit on the complexity of the questionnaire to avoid confusion or errors, compared to interviewers trained in questionnaire use. This method is not appropriate for potential respondents with limited ability to read or write English unless questionnaires are provided in the appropriate language.

### Dropoff-Mailback and Dropoff-Pickup

These two closely-related methods of self-enumeration surveys both provide higher response rates than postal surveys. In both cases, the questionnaire is delivered to respondents by an interviewer who explains the aims of the survey and how to fill out the questionnaire. The questionnaire is left with the respondent to be completed and either mailed back by the respondent or picked up by the interviewer at a later date. The costs involved in using interviewers makes this technique more expensive than postal surveys, but usually less expensive than personal interviews. This method has the same disadvantage as personal interviewing in that the respondent has to be available when the interviewer visits.

### Electronic Form or CASI

An electronic form (eform) is an electronic version of the questionnaire that can be sent to the respondent's computer via email or accessed from the World Wide Web. The data entered onto an eform can be edited as it is entered, thus improving data quality. The use of electronic returns also produces a faster response than other self-enumeration methods. Questions can also be sequenced so that only the questions relevant to the respondent are visible. The disadvantages are: increased cost for development of the forms, maintenance of the related systems and security, and the absence of well-established design standards. Also, this technique requires respondents to have compatible computer software and help-desk type staff can be necessary to support the use of the form. Electronic forms on disk or CD are also sometimes used in mail-out or drop-off surveys instead of a paper form.

## TELEPHONE INTERVIEWS

### Telephone Interview

In the ABS telephone surveys are used for both household and business surveys to some extent, however telephone data collection methods are used more widely for intensive follow-up and post-enumeration tests. Telephone interviews involve potential respondents being telephoned and asked the survey questions over the phone. This technique:

- reduces costs, compared to personal interviews, as fewer interviewers are needed and there are no travel costs involved,
- produces more timely results, and
- call-backs for people 'not-answering' and follow-ups for additional information are relatively quick and inexpensive.

However, telephone interviews are very limited in the number and complexity of the questions that can be asked and, because of the ease with which the respondent can terminate the interview, non-response and partial non-response can be high. A bias is also introduced because people with no phones or who are rarely available are excluded and it is difficult to produce a reliable sampling frame because of unlisted numbers and changes in addresses (therefore changes in numbers). There are also problems with convincing respondents of the authority behind the survey and the confidentiality of results.

### Computer-Assisted Telephone Interview

Computer assisted data collection methods can also be used in telephone surveys. Computer Assisted Telephone Interviewing (CATI) involves responses being keyed directly into a computer by the telephone interviewer. This technique allows for:

- some editing to be carried out immediately (which improves the data quality and decreases processing time),
- 'call scheduling' to take place. Respondents can be called at convenient times or when data is available. Also, if the phone is engaged, the system will reschedule the call, and
- questions to be sequenced so that only relevant questions are visible to the interviewer (therefore reducing interviewer errors).
- monitoring of interviewing staff so that consistency of performance is higher.

However, CATI is expensive to implement and maintain with extensive training needed for interviewers to perform properly. Interview time is also increased because editing is done whilst the interview is being performed and this can cause increased non-response.

This page last updated 1 March 2023

## FRAMES & POPULATION

### Introduction

Once the aims of the survey, data requirements, level of accuracy, data collection method and costs have been worked out, the sample design process is next considered. The sample design refers to what a sample consists of and how the sample is to be obtained. It is concerned with defining the population and the frame, sample size issues, sampling techniques and data collection method. In this chapter, frames and population will be discussed and the other issues will be discussed later in other chapters.

### Population

The population is the aggregate or collection of units about which the survey will be conducted. Units can refer to people, households, schools, hospitals, businesses etc. There are two different populations that a survey is concerned with. We have a target population, the group of units about which information is wanted, and a survey population, the units that we are able to survey. The target population is also known as the *scope* of the survey, the population that the survey is aimed at; the survey population is also called the coverage, the population the survey actually covers. Ideally the survey population should correspond exactly with the target population, however, the two populations may not match, so the conclusions based on survey data only apply to the survey population.

### Frame

The frame refers to the list of units (eg, persons, households, businesses, etc) in the survey population. Since the selection of the sample is directly based on this list, the frame is one of the most important tools in the design of a survey. It determines how well a target population is covered, and affects the choice of the data collection method. It is also desirable that the frame contains auxiliary information on the units so that a more efficient sample plan can be developed (auxiliary information is discussed in Sample Design). The frame should contain contact points for each of the units listed so that it can be used to access the population. This means that for postal surveys the frame should contain postal addresses; for interviewer-based surveys the frame should contain street addresses; and for telephone surveys the frame should contain telephone numbers.

## TYPES OF FRAMES

There are two types of frames that may be used for survey design: list frames and area frames.

### List Frames

A list frame is a list of all of the selection units in the survey population. List frames are commonly used in surveys of businesses. Examples of list frames include administrative lists, personnel lists, telephone lists, mailing house lists, association membership lists and the electoral roll.

### Area Frames

An area frame is a complete and exhaustive list of non-overlapping geographic areas. These areas may be defined by such geographic features as rivers and streets. Area frames are used when it is too expensive or complex to maintain a list frame or where no list frame exists and it would be too expensive or complex to create. They are usually used in household surveys where lists of households are only created for selected geographic regions. Examples of the geographic areas that may be used to create an area frame include Local Government Areas (LGAs), Census Collector's District (CDs), Postcodes, and States.

## FRAME PROBLEMS AND SOLUTIONS

For most sampling methodologies it is desirable to have a complete list of units or areas from which to select a sample. However, in practice it can be difficult to compile such a complete list and therefore frame bias (ie. the frame is not representative of the target population) is introduced. The bias might result for two reasons: use of the inappropriate frame which may include some out of scope units or may exclude some appropriate units (eg. choosing the wrong units from the frame, for example, choosing non-retail units from a list of businesses when conducting a survey of retail sales); and problems with the composition of the frame

Frames can become inaccurate for many reasons:

- the most common being that populations are subject to continuous change and the frame easily becomes out of date; also
- frames are often compiled from inadequate sources, this can sometimes cause frame units to be hard to contact through lack of information.

Some of the problems that can occur in the composition of the frame are described below.

**Missing Units**
Missing units (eg exporting businesses) on a frame are those units in the target population that should appear on the frame but do not. These units may have different characteristics to those units which do appear on the frame and therefore, information obtained from the survey will not be representative of the target population. This is referred to as under-coverage and may result in bias.

**Out -of -Scope (or Foreign) Units**
These are units that do appear on the frame but are not part of the target population (eg non-exporting businesses). These units do not contribute to the survey results but they do contribute to costs. Selection of a number of foreign elements in the sample reduces the actual sample size, causing larger errors in survey estimates.

**Duplications**
Duplications refer to units that appear on the frame more than once. This can be due to;

- typographical errors in entering units onto the frame;
- the same unit is entered under slightly different names;
- merging of several smaller frames when the frame is created.

This means that the probability of selection of the units on the frame is no longer known. A unit may be selected more than once in the sample and this reduces the accuracy of the results because fewer different units are sampled. It can also reduce the professional image of the survey.

**Deaths**
A dead unit is a unit that no longer exists in the population (eg exporting business has folded and no longer exists). Deaths on the frame have a similar effect on survey results as out-of-scope units. Deaths should not be removed from the frame for future surveys based on sample survey results as deaths in the sample reflect the number of deaths in the non-sample population. Retaining dead units in the sample provides an indication of the number of dead units in the survey population. Deaths, however, can be removed based on census results.

**Nils**
Whilst nils are not a frame problem they are worth mentioning here as they can often be confused with dead units. Units that are operating but have a zero return or activity for the survey period are referred to as nils. The zero return for the survey period may be due to a seasonal factor inherent to the business (eg. beachside ice-cream stalls would report zero turnover during winter). These units should not be removed from the frame as they will report non-zero returns at a later date.

**Solutions to Frame Problems**
It is important to be aware that frames do have problems, so the quality of the frame should be investigated. There are various strategies that may be used if the quality of the frame is in doubt:

- You can use the frame anyway and allow for the problems by increasing the sample size at the selection stage and by adjusting the weights at the estimation stage.
- If time and resources allow, it may be possible to update the frame.
- If another frame exists that also closely approximates the target population, it may be better to use the alternative frame. This situation calls for a trade-off between a frame matching the target population and a frame not quite matching the target population but providing the relevant detailed information.
- It may also be possible to combine the frame with related frames to improve the coverage of the target population. This process is often called *supplementation* as the current frame is supplemented with another one. However, consideration should be given to overlapping units and any differences between the definition of a unit on the two frames.

## CONCLUSION

Since the frame provides the means of accessing the population to obtain a sample, considerations should therefore be given to the quality of the frame. Frames should be evaluated early in the planning stage since a bad frame will have an affect on the estimates produced at the end.

A good frame is up-to-date, does not have any missing units, contains only relevant units, does not include duplicates, is accessible to frame users and contains sufficient information to uniquely identify and contact each unit.

This page last updated 1 March 2023

## ERRORS IN STATISTICAL DATA

### Introduction

The accuracy of a survey estimate refers to the closeness of the estimate to the true population value. Where there is a discrepancy between the value of the survey estimate and true population value, the difference between the two is referred to as the error of the survey estimate. The total error of the survey estimate results from the two types of error:

- sampling error, which arises when only a part of the population is used to represent the whole population; and
- non-sampling error which can occur at any stage of a sample survey and can also occur with censuses. Sampling error can be measured mathematically whereas measuring non-sampling error can be difficult.

It is important for a researcher to be aware of these errors, in particular non-sampling error, so that they can be either minimised or eliminated from the survey. An introduction to measuring sampling error and the effects of non-sampling error is provided in the following sections.

### Sampling Error

Sampling error reflects the difference between an estimate derived from a sample survey and the "true value" that would be obtained if the whole survey population were enumerated. It can be measured from the population values, but as these are unknown (otherwise there would be no need for a survey), it can also be estimated from the sample data. It is important to consider sampling error when publishing survey results as it gives an indication of the accuracy of the estimate and therefore reflects the importance that can be placed on interpretations. If sampling principles are applied carefully within the constraints of available resources, sampling error can be accurately measured and kept to a minimum.

### Factors Affecting Sampling Error

Sampling error is affected by a number of factors including sample size, sample design, the sampling fraction and the variability within the population. In general, larger sample sizes decrease the sampling error, however this decrease is not directly proportional. As a rough rule of thumb, you need to increase the sample size fourfold to halve the sampling error. Of much lesser influence is the sampling fraction (the fraction of the population size in the sample), but as the sample size increases as a fraction of the population, the sampling error should decrease.

The population variability also affects the sampling error. More variable populations give rise to larger errors as the samples or the estimates calculated from different samples are more likely to have greater variation. The effect of the variability within the population can be reduced by increasing the sample size to make it more representative of the survey population. Various sample design options also affect the size of the sampling error. For example, stratification reduces sampling error whereas cluster sampling tends to increase it (these designs are discussed in Sample Design).

### Standard Error

The most commonly used measure of sampling error is called the standard error (SE). The standard error is a measure of the spread of estimates around the "true value". In practice, only one estimate is available, so the standard error can not be calculated directly. However, if the population variance is known the standard error can be derived mathematically. Even if the population variance is unknown, as happens in practice, the standard error can be estimated by using the variance of the sample units. Any estimate derived from a probability based sample survey has a standard error associated with it (called the standard error of the estimate, written se(y) where y is the estimate of the variable of interest). Note that :

- The standard error is an indication of how close the sample survey estimate is to the result that would have been obtained from a census under the same operating conditions (an equal complete coverage).
- The standard error only gives a measure of the variation in values obtained from repeated samples. It does not measure the precision of the particular sample from which it is estimated.
- A small standard error indicates that the variation in values from repeated samples is small and therefore the chance of a 'bad' sample is small - hence there is more likelihood that the sample estimate will be close to the result of an equal complete coverage.
- Standard errors can be used to work out upper and lower limits ('confidence interval'), which will include the result from an equal complete coverage with a certain probability.
- Estimates of the standard error can be obtained from any one of the possible random samples.

- The standard error calculated from a sample is itself an estimate (and is also subject to sampling error)
- When publishing the results of any survey, statements about the standard error of the estimates should be made.
- When comparing survey estimates, the standard errors must be taken into account.
- The term 'sampling variance' refers to the square of the standard error.

For more information on how to calculate estimates and their standard errors please refer to Analysis.

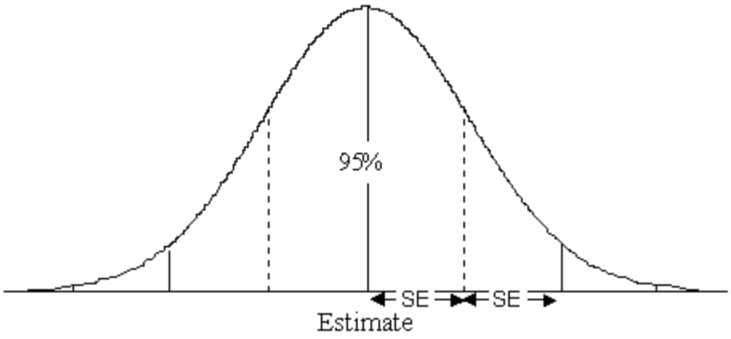
**Variance**
The variance is another measure of sampling error, which is simply the square of the standard error:
$\mathrm{Var}(y) = \mathrm{se}(y)^2$

**Relative Standard Error**
Another way of measuring sampling error is the relative standard error (RSE) where the standard error is expressed as a percentage of the estimate. The RSE avoids the need to refer to the estimate and is useful when comparing variability of population estimates with different means. RSE is an important measure when expressing the magnitude of standard error relative to the estimate. The relative standard error is calculated as follows (where y is the estimate of the variable of interest):

$\mathrm{RSE}(y) = 100 * \{\mathrm{se}(y) / y\}$

**Confidence Interval**
Assuming that the target population is distributed normally for the characteristic being measured, (or, if estimating the mean, the sample is sufficient to assume the sample mean is distributed normally) the interval which contains the true value is usually calculated as being one, two, or three standard errors above and below the survey estimate. This interval is usually referred to as a confidence interval.

Normal Curve



There is a 95% chance that the confidence interval which extends to two standard errors on either side of the estimate contains the "true value". This interval is called the 95% confidence interval and is the most commonly used confidence interval. The 95% confidence interval is written as follows:

$95\%\ \mathrm{CI}(y) = [y - \{2*\mathrm{se}(y)\} , y + \{2*\mathrm{se}(y)\}]$

This is expressed: "We are 95% confident that the true value of the variable of interest lies within the interval [y - {2*se(y)} , y + {2*se(y)}]".

Other confidence intervals are the 68% confidence interval (where the confidence interval extends to one standard error on either side of the estimate has a 68% chance of containing the "true value") and the 99% confidence interval (where the confidence interval extends to three standard errors on either side of the survey estimate has a 99% chance of containing the "true value").

For example, suppose a survey estimate is 50 with a standard error of 10. The confidence interval 40 to 60 has a 68% chance of containing the "true value", the interval 30 to 70 has a 95% chance of containing the "true value" and the interval 20 to 80 has a 99% chance of containing the "true value".

**NON-SAMPLING ERROR**
Non-sampling error is all other errors in the estimate. Some examples of causes of non-sampling error are non-response, a badly designed questionnaire, respondent bias and processing errors.

Non-sampling errors can occur at any stage of the process. They can happen in censuses and sample surveys. Non-sampling errors can be grouped into two main types: systematic and variable.

Systematic error (called bias) makes survey results unrepresentative of the target population by distorting the survey estimates in one direction. For example, if the target population is the population of Australia but the survey population is just males then the survey results will not be representative of the target population due to systematic bias in the survey frame.

Variable error can distort the results on any given occasion but tends to balance out on average. Some of the types of non-sampling error are outlined below:

### Failure to Identify Target Population / Inadequate Survey Population

The target population may not be clearly defined through the use of imprecise definitions or concepts. The survey population may not reflect the target population due to an inadequate sampling frame and poor coverage rules. Problems with the frame include missing units, deaths, out-of-scope units and duplicates. These are discussed in detail in Frames and Population.

### Non-Response Bias

Non-respondents may differ from respondents in relation to the attributes/variables being measured. Non-response can be total (none of the questions answered) or partial (some questions may be unanswered owing to memory problems, inability to answer, etc.). To improve response rates, care should be taken in designing the questionnaires, training of interviewers, assuring the respondent of confidentiality, motivating him/her to co-operate, and calling back at different times if having difficulties contacting the respondent. "Call-backs" are successful in reducing non-response but can be expensive for personal interviews. Non-response is covered in more detail in Non-Response.

### Questionnaire problems

The content and wording of the questionnaire may be misleading and the layout of the questionnaire may make it difficult to accurately record responses. Questions should not be loaded, double-barrelled, misleading or ambiguous, and should be directly relevant to the objectives of the survey.

It is essential that questionnaires are tested on a sample of respondents before they are finalised to identify questionnaire flow and question wording problems, and allow sufficient time for improvements to be made to the questionnaire. The questionnaire should then be re-tested to ensure changes made do not introduce other problems. This is discussed in more detail in Questionnaire Design.

### Respondent Bias

Refusals to answer questions, memory biases and inaccurate information because respondents believe they are protecting their personal interest and integrity may lead to a bias in the estimates. The way the respondent interprets the questionnaire and the wording of the answer the respondent gives can also cause inaccuracies. When designing the survey you should remember that uppermost in the respondent's mind will be protecting their own personal privacy, integrity and interests. Careful questionnaire design and effective questionnaire testing can overcome these problems to some extent. Respondent bias is covered in more detail below.

### Processing Errors

There are four stages in the processing of the data where errors may occur: data grooming, data capture, editing and estimation. Data grooming involves preliminary checking before entering the data onto the processing system in the capture stage. Inadequate checking and quality management at this stage can introduce data loss (where data is not entered into the system) and data duplication (where the same data is entered into the system more than once). Inappropriate edit checks and inaccurate weights in the estimation procedure can also introduce errors to the data. To minimise these errors, processing staff should be given adequate training and realistic workloads.

### Misinterpretation of Results

This can occur if the researcher is not aware of certain factors that influence the characteristics under investigation. A researcher or any other user not involved in the collection stage of the data gathering may be unaware of trends built into the data due to the nature of the collection, such as it's scope. (eg. a survey which collected income as a data item with the survey coverage and scope of all adult persons (ie. 18 years or older), would expect to produce a different estimate than that produced by the ABS Survey of Average Weekly Earnings (AWE) simply because AWE includes persons of age 16 and 17 years as part of it's scope). Researchers should carefully investigate the methodology used in any given survey.

### Time Period Bias

This occurs when a survey is conducted during an unrepresentative time period. For example, if a survey aims to collect details on ice-cream sales, but only collects a weeks worth of data during the hottest part of summer, it is unlikely to represent the average weekly sales of ice-cream for the year.

### Minimising Non-Sampling Error

Non-sampling error can be difficult to measure accurately, but it can be minimised by

- careful selection of the time the survey is conducted,
- using an up-to-date and accurate sampling frame,
- planning for follow up of non-respondents,
- careful questionnaire design,
- providing thorough training for interviewers and processing staff and
- being aware of all the factors affecting the topic under consideration.

## RESPONDENT BIAS

No matter how good the questionnaire or the interviewers are, errors can be introduced into a survey either consciously or unconsciously by the respondents. The main sources of error relating to respondents are outlined below.

### Sensitivity

If respondents are faced with a question that they find embarrassing, they may refuse to answer, or choose a response which prevents them from having to continue with the questions. For example, if asked the question: "Are you taking any oral contraceptive pills for any reason?", and knowing that if they say "Yes" they will be asked for more details, respondents who are embarrassed by the question are likely to answer "No", even if this is incorrect.

### Fatigue

Fatigue can be a problem in surveys which require a high level of commitment from respondents. For example, diary surveys where respondents have to record all expenses made in a two week period. In these type of surveys, the level of accuracy and detail supplied may decrease as respondents become tired of recording all expenditures.

## NON-RESPONSE

Non-Response results when data is not collected from respondents. The proportion of these non-respondents in the sample is called the *non-response rate*. Non-response can be either partial or total. It is important to make all reasonable efforts to maximise the response rate as non-respondents may have differing characteristics to respondents. This causes bias in the results.

### Partial Non-Response

When a respondent replies to the survey answering some but not all questions then it is called partial non-response. Partial non-response can arise due to memory problems, inadequate information or an inability to answer a particular question. The respondent may also refuse to answer questions if they

- find questions particularly sensitive, or
- have been asked too many questions (the questionnaire is too long).

### Total Non-Response

Total non-response can arise if a respondent cannot be contacted (the frame contains inaccurate or out-of-date contact information or the respondent is not at home), is unable to respond (may be due to language difficulties or illness) or refuses to answer any questions.

When conducting surveys it is important to collect information on why a respondent has not responded. For example when evaluating a program a respondent may indicate they were not happy with the program and therefore do not wish to be part of the survey. Another respondent may indicate that they simply don't have the time to complete the interview or survey form. If a large number of those not responding indicate dissatisfaction with the program, and this is not indicated in the final report, an obvious bias would be introduced in the results.

### Minimising Non-Response

Response rates can be improved through good survey design via short, simple questions, good forms design techniques and explaining survey purposes and uses. Assurances of confidentiality are very important as many respondents are unwilling to respond due to a fear of lack of privacy. Targeted follow-ups on non-contacts or those initially unable to reply can increase response rates significantly.

Following are some hints on how to minimise refusals in a personal or phone contact:

Find out the reasons for refusal and try to talk through them

- Use positive language
- State how and what you plan to do to help with the questionnaire
- Stress the importance of the survey
- Explain the importance of their response as a representative of other units
- Emphasise the benefits from the survey results, explain how they can obtain results
- Give assurance of the confidentiality of the responses

Other measures that can improve respondent cooperation and maximise response include:

- Public awareness activities including discussions with key organisations and interest groups, news releases, media interview and articles. This is aimed at informing community about the survey, identifying issues of concern and addressing them.
- Advice to selected units by letter, giving them advance notice and explaining the purposes of the survey and how the survey is going to be conducted.

In case of a mail survey most of the points above can be stated in an introductory letter or through a publicity campaign.

### Allowing for Non-Response

Where response rates are still low after all reasonable attempts of follow-up are undertaken, you can reduce bias by using population benchmarks to post-stratify the sample (covered in Sample Design), intensive follow-up of a subsample of the non-respondents or imputation for item non-response (non-response to a particular question).

The main aim of imputation is to produce consistent data without going back to the respondent for the correct values thus reducing both respondent burden and costs associated with the survey. Broadly speaking the imputation methods fall into three groups:

- the imputed value is derived from other information supplied by the unit;
- values by other units can be used to derive a value for the non-respondent (eg average);
- an exact value of another unit (called donor) is used as a value for the non-respondent (called recipient);

When deciding on the method of imputation it is desirable to know what effect will imputation have on the final estimates. If a large amount of imputation is performed the results can be misleading, particularly if the imputation used distorts the distribution of data.

If at the planning stage it is believed that there is likely to be a high non-response rate, then the sample size could be increased to allow for this. However, the non-response bias will not be overcome by just increasing the sample size, particularly if the non-responding units have different characteristics to the responding units. Post-stratification and imputation also fail to totally eliminate non-response bias from the results.

### Example: Effect of Non-Response

Suppose a postal survey of 3421 fruit growers was run to estimate the average number of fruit trees on a farm. There was an initial period for response and following low response rates, two series of follow up reminders were sent out. The response and results were as follows:

| | Response | Ave. no. of Trees |
|---|---:|---:|
| Initial Response | 300 | 456 |
| Added after 1 follow up reminder | 543 | 382 |
| Added after 2 follow up reminders | 434 | 340 |
| Total Response | 1277 | |

After two follow up reminders there was still only a 37% response rate. From other information it was known that the overall average was 329. The result based on this survey would have been:

| | Cumulative Response | Combined Average |
|---|---:|---:|
| Initial Response | 300 | 456 |

| | | |
|---|---|---|
| Added after 1 follow up reminder | 843 | 408 |
| Added after 2 follow up reminders | 1277 | 385 |

If results had been published without any follow-up then the average number of trees would have been too high as farms with greater number of trees appeared to have responded more readily. With follow-up, more smaller farms sent back survey forms and the estimate became closer to the true value.

This page last updated 1 March 2023

## SAMPLE DESIGN

### Introduction

When you have a clear idea of the aims of the survey, the particular data requirements, the degree of accuracy required, and have considered the resources and time available, you are in a position to make a decision on the size and form of the collection. The major concerns that should be addressed at this stage are:

- defining the population, frame and units;
- calculating the sample size;
- determining the sampling methodology;
- choosing an appropriate data collection method; and
- determining the estimation method to be used.

The following discussion will give a brief introduction to some basic terms and ideas in sampling and an outline of sample designs commonly used. The main focus of the discussion will be on determining an appropriate sampling method.

## PROBABILITY AND NON-PROBABILITY

### Non-Probability Samples

If the probability of selection for each unit is unknown, or cannot be calculated, the sample is called a non-probability sample. Non-probability samples are often less expensive, easier to run and don't require a frame.

However, it is **not** possible to accurately evaluate the precision (ie. closeness of estimates under repeated sampling of the same size) of estimates from non-probability samples since there is no control over the representativeness of the sample. If a non-probability sample is carried out carefully, then the bias in the results can be reduced.

As it is dangerous to make inferences about the target population on the basis of a non-probability sample, non-probability methodology is often used to test aspects of a survey such as questionnaire design, processing systems etc. rather than make inferences about the target population.

Different types of non-probability samples are discussed below.

### Quota Sampling

To select a quota sample, the interviewers select respondents until a pre-determined number of respondents in certain categories are surveyed (eg. the interviewers might select the sample to achieve a certain age/sex breakdown reflective of the target population).

This is the method of sampling commonly used by market researchers and political pollsters as it can produce fairly good estimates if it is properly conducted. When top up units are selected randomly to fill a quota, and no element of judgment is used by the researcher for unit selection, it is very similar to a probability sample. However, when non-response is significant (which is almost always the case for voluntary surveys), quota sampling can under-represent those portions of the population that are unwilling to respond or hard to contact. This is of particular concern when the data items collected influence the likelihood of response. See also the section on Non-Response in Errors in Statistical Data for further details.

### Convenience and Haphazard Sampling

Street corner interviews, magazine and newspaper questionnaires and phone-in polls are all examples of convenience or haphazard samples. These types of surveys are subject to biased or unrepresentative samples as only persons who feel strongly about the topic will respond. These surveys also have a tendency to ask questions that are loaded or have a biased wording.

Street corner interviews can be biased depending on the timing and the placement of the interviewer. There is no control over selecting the sample of respondents in any of these methods, however they are very cheap and easy to administer.

### Judgement or Purposive Sampling

Judgement sampling is where a 'representative' sample is chosen by an expert in the field of study. Judgement sampling is subject to unknown biases but may be justified for very small samples. This form of sampling can be used to choose a sample for a pilot test of a probability survey but inferences about the population should not be made from judgement samples. Judgement sampling is also known as purposive sampling.

### Probability Samples

A probability sample is one in which every unit of the population has a known non-zero probability of selection and is randomly selected.

A probability sample allows inferences about the target population to be made. By knowing the selection probability for each unit, objective selections can then be made which should produce a more representative sample. Known probabilities also allow the measurement of the precision of the survey estimates in terms of standard errors and confidence intervals.

Probability samples require a frame for selection purposes and thus are relatively expensive in terms of operational costs and frame maintenance. The most common sampling techniques, such as simple random, systematic, stratified, multi-stage and cluster sampling, are all examples of probability samples. These will be looked at later in this chapter.

### Choosing Between Probability and Non-Probability Samples

The choice between using a probability or a non-probability approach to sampling depends on a variety of factors:

- the objectives and scope of the survey;
- the method of data collection suitable to those objectives;
- the precision required of the results and whether that precision needs to be able to be measured;
- the availability of a sampling frame;
- the resources required to maintain the frame; and
- the availability of extra information about the units in the population.

Probability sampling is normally preferred when conducting major surveys, especially when a population frame is available ensuring that we are able to select and contact each unit in the (frame) population. However, where time and financial constraints make probability sampling infeasible, or where knowing the level of accuracy in the results is not an important consideration, non-probability samples do have a role to play since they are inexpensive, easy to run and no frame is required. For this reason, when conducting qualitative (investigative), rather than quantitative research, non-probability samples & techniques such as case studies are generally superior to probability samples & quantitative estimation. Non-probability sampling can also be useful when pilot testing surveys.

If a non-probability sample is carried out carefully, then the bias in the results can be reduced. Note that with non-probability methods it is dangerous to make inferences about the whole population.

Quota sampling may be appropriate when response rates are expected to be low. True probability sampling would be more expensive and may require top up units to be selected. If quota sampling is used, selection of units should be as random as possible and care should be taken to avoid introducing a bias.

Unlike certain non-probability samples, probability sampling involves a random selection of units. This allows us to quantify the standard error of estimates and hence allow confidence intervals to be formed and hypotheses to be formally tested. The main disadvantages with probability sampling involve cost, such as the costs involved with frame maintenance and surveying units which are difficult to contact.

## SIMPLE RANDOM SAMPLING

Simple random sampling (SRS) is a probability selection scheme where each unit in the population is given an equal probability of selection, and thus every possible sample of a given size has the same probability of being selected. One possible method of selecting a simple random sample is to number each unit on the sampling frame sequentially, and make the selections by generating "selection numbers" from a random number table or, from some form of random number generator.

### With (SRSWR) and Without (SRSWOR) Replacement

Simple random sampling can involve the units being selected either with or without replacement. With replacement sampling allows the units to be selected multiple times whilst without replacement only allows a unit to be selected once. Without replacement sampling is by far the more commonly used method.

### Advantages

The advantage of simple random sampling lies in its simplicity and ease of use, especially when only a small sample is taken.

### Disadvantages

Simple random sampling does, however, require a complete list of all population units as each unit needs to have a unique number associated with it to enable random selection. This sampling scheme also becomes unwieldy for large sample sizes and can be expensive if the sample is spread over a wide geographic area.

In practice, simple random sampling is rarely used because there is almost always a more efficient method of designing the sample (in terms of producing accurate results for a given cost). Nevertheless, simple random sampling forms the basis of a number of the more complex methods of sample design, and is used as a benchmark to which other designs are compared.

## USE OF AUXILIARY INFORMATION

We have discussed methods of drawing simple random samples.

There is no point in using any other kind of sample selection method if one knows no more about the population to be sampled than the existence of each of the units in the population. However, some further information is often available about each of the population units from simple observation or data from a previous study (for example, a census). This information can be in the form of demographic variables such as age, sex and income or geographical/business types such as industry, employment, state, region and sector (private or public).

This further information about the population units is called auxiliary information. Such information can be used in the selection and the estimation process to obtain more accurate estimates or reduce costs. Sampling techniques using this information include systematic sampling, stratified sampling (including post-stratification) and cluster sampling. Auxiliary information is also used in estimation techniques such as ratio and regression estimations.

One of the major aspects of sample design is the efficient use of auxiliary or supplementary information. As there is often a cost involved in obtaining auxiliary information, it is necessary to quantify the gains that are obtained through using auxiliary information and balance it against the cost of acquisition.

## SYSTEMATIC SAMPLING

Systematic sampling provides a simple method of selecting the sample when the sampling frame exists in the form of an explicit list. Where the frame contains auxiliary information then the units in the frame are ordered with respect to that auxiliary data (eg employment size of a business). A fixed interval (referred to as the skip) is then used to select units from the sampling frame. Systematic sampling is best explained by describing how the sample selections are made.

### Method

Assume that the population has N units (eg. 37) and that we wish to select a sample of size n (eg. 5). The list of N population units is ordered in some way. The steps taken in selecting a systematic random sample are:

1. Calculate the skip interval k = N/n.
2. Choose a random start, r, between 1 and k.
3. Select the rth unit in the list and every kth unit thereafter:

r, r+k, r+2*k, r+3*k,..., r+(n-1)*k.

The value of k is usually not an integer. In this case we either

- round k to the nearest integer;
- keep k a fraction and round r+i*k (where i =0,...,n-1) to the nearest integer.

### Example: Calculating the Skip Interval

Say that we wanted to take a systematic sample of size 5 from a population of 37 units. The sample size does not divide evenly into the population. The two options for coping with this are discussed below.

Order the population units in some way and number them from 1 to 37.

N = 37
n = 5
k = 37/5 = 7.4

1. Round k to the nearest integer

k = 7
r = 4

Then the sample units are :
1st unit=4, 2nd unit=4+7=11, 3rd unit=4+14=18, 4th unit=4+21=25 and 5th unit=4+28=32.
Therefore, sample = ( 4, 11, 18, 25, 32)

2. Round r+ik to the nearest integer

k = 7.4
r = 4.2
sample = ( 4.2, 11.6, 19, 26.4, 33.8) = ( 4, 12, 19, 26, 34)

**Features of Systematic Random Sampling**

The usefulness of systematic random sampling depends upon the strength of the relationship between the variable of interest and the benchmark variable/s. The more highly correlated they are, the greater the gains in accuracy achieved over simple random sampling. This is because we are ensuring a more representative sample of population units are selected. If there is a strong relationship between the variable of interest and the benchmark variable/s then ordering the list by the variable of interest will yield more accurate results using systematic sampling than simple random sampling. Systematic random sampling using ordered lists ensures a range of units will be selected in the sample.

**Advantages**

- Ease of use

  Systematic random sampling is often easier to use than simple random sampling, especially for large samples, as only one random number (the random start) is required, rather than a random number for every unit as required for SRS.

- More accurate results

  Systematic random sampling is a without replacement sampling scheme and usually gives more accurate results (lower standard errors) than simple random samples of the same size due to the closer control over selection. This is particularly the case if the ordering of the units in the list is related to characteristics of the variable of interest.

**Disadvantages**

- Periodicity bias

  If, after ordering, the variable of interest is periodic/cyclic in nature then it is possible to obtain an estimate which is less accurate than a simple random sample if the periodicity coincides with the skip interval.

  For example, daily sales in a supermarket expect to peak on weekends. If the skip is calculated as 7, a bias is introduced yielding samples which are not representative of the population.
- Complete list of the population

  To perform the ordering, a complete list of the population is required.

**Stratified Sampling**

Stratified sampling is a technique which uses auxiliary information which is referred to as *stratification* *variables* to increase the efficiency of a sample design. Stratification variables may be geographical (eg. state, rural/urban) or non-geographical (eg. age, sex, number of employees).

Stratified sampling involves

- the division or stratification of the population into homogeneous (similar) groups called strata; and
- selecting the sample using SRS or systematic sampling within each stratum and independent of the other strata.

Stratification almost always improves the accuracy of estimates. This is because the population variability can be thought of as having components within strata and between strata. By independently sampling within each stratum we ensure each stratum is appropriately reflected in the sample, so between stratum variability is eliminated and we are left only with the within stratum component. With this factor in mind we see that the most efficient way to stratify is to have strata which are as different from each other as possible (to maximise the variance which is being eliminated) while being internally as homogeneous as possible (to minimise the variance remaining).

### Practical Considerations

When planning a stratified sample, a number of practical considerations should be kept in mind:

- the strata should be designed so that they collectively include all members of the target population;
- each member must appear in only one stratum, ie strata should be non overlapping; and
- the definitions of boundaries of the strata should be precise and unambiguous.

### Example of Stratification

As an example of stratification, if we were interested in the educational background of members of a Science faculty at a University, we could select a sample from the faculty as a whole or select samples independently from each of the departments within the faculty, such as mathematics, physics, chemistry etc. This latter method would ensure that each department was adequately represented (which would not necessarily happen otherwise), and should increase the precision of the overall estimate.

If on the other hand, we were interested in the level of education (PhD, Masters, Bachelor) rather than the background we should stratify the faculty by level (Professor, Senior Lecturer, Lecturer) rather than by the department. Using this stratification we are more likely to find uniformity of educational standards within a level rather than an area of work, and we are also more likely to separate the better qualified from the less qualified.

### Advantages

The four main benefits of stratified sampling are:

- minority groups of interest or high variability can be oversampled. A greater proportion of units can be selected from minority groups than the majority group.
- the results are more accurate. Sampling error is reduced because of the grouping of similar units. It should be remembered that there is no gain in accuracy from stratifying by a factor unrelated to the subject of the survey.
- different selection and estimation procedures can be applied to the various strata; and
- separate information can be obtained about the various strata. Stratification also permits separate analyses on each group and allows different interests to be analysed for different groups.

### Disadvantages

- an increase in costs;
- a danger of stratifying too finely;
- availability of auxiliary information (however, it is possible to form a stratum of units for which no information is available)

### Number of Strata

There is no rule as to how many strata the population should be divided into. This depends on the population size and homogeneity and the format in which the output is required. If output is required for some sub-groups of the population these subgroups must be considered as separate strata.

### ABS Surveys

All surveys conducted by the Australian Bureau of Statistics employ stratification. Household surveys (such as the Monthly Population Survey and the Household Expenditure Survey) use geographic strata. Business surveys use variables such as state and industry strata and use some measure of size (eg employment) to form size strata.

### Allocation of Sample

An important consideration after deciding on the appropriate stratification is the way in which the total sample is to be allocated to each stratum. There are three common methods of calculating the number of units required from each stratum.

- Equal Allocation
  Equal allocation is the simplest form of allocation. It involves selecting the same number of units from each stratum. This method is rarely used as it does not take into account the size of each stratum or the variability within each strata. However, equal allocation does ensure approximately equal reliability across all strata (assuming that the sample is not a significant proportion of the population [no more than 10% of the population would be reasonable]).

- Proportional Allocation
  In proportional allocation, the sample allocated to each stratum is proportional to the number of units in the strata. Say we were sampling 10% of the population; we would then sample 10% of each stratum. This method takes into account the size of each stratum; larger strata will have larger samples taken from them. It does not however allow for the differences in the variability within each strata. This method is used when no information is available on the variability within the stratum (stratum variances) or where there is no major difference in variability between the strata.

- Optimal Allocation
  This allocation method aims to minimise the standard error of the estimate across the population. This is achieved by allocating a relatively large sample to those strata which are highly variable, and a relatively small sample to those strata which have low variability. Variability is based on some auxiliary information available for the stratum, for example, previous survey results or pilot test data. Optimal allocation can also be used to account for differing costs of sampling between the strata, ie. minimise the standard error for a given cost.

Another method of sample selection is to have a completely enumerated stratum in our sample. This is where the units that contribute significantly to our estimates are placed in a single stratum and every unit within the stratum is then selected.

### Estimation

As samples are selected independently from each stratum, estimates are also usually made separately for each stratum, then added to give the overall estimate (eg estimated unemployment for Australia will be the sum of the state unemployment estimates).

Similarly, standard errors or variances (measures of sample variability) are calculated for each stratum and then all strata specific variances are added up to obtain the overall variance. The addition of variances is possible because the sample is selected independently from each strata. This overall variance can then be used to calculate an overall standard error.

### Post Stratification

There will be occasions when we may like to stratify by a certain variable, say age or sex, but we cannot because we do not know the age and sex of our population units until we select them.

Post-stratification is a method used when stratification is not possible before the survey. The stratification variable can then be used after the survey is conducted, to improve the efficiency of estimates or, to obtain estimates corresponding to different categories of that variable (eg. sex) by stratifying the sample as if the benchmark information had been available previously.

## CLUSTER AND MULTI-STAGE SAMPLING

So far we have considered a number of ways which a sample of population units can be selected and population characteristics estimated on the basis of this sample. In this section consideration is given to a sampling scheme where the selection of population units is made by selecting particular groups (or clusters) of such units and then selecting all or some of the population units within selected groups for inclusion in the sample.

### Cluster Sampling

Cluster sampling involves selecting a sample in a number of stages (usually two). The units in the population are grouped into convenient, usually naturally occurring clusters. These clusters are non-overlapping, well-defined groups which usually represent geographic areas. At the first stage of selection, a number of clusters are selected. At the second stage, all the units in the chosen clusters are selected to form the sample.

### Practical Considerations

- The clusters should be designed so that they collectively include all members of the target population;
- each member must appear in one and only one cluster; and
- the definitions or boundaries of the clusters should be precise and unambiguous; in the case of geographical clusters natural and man-made boundaries such as rivers and roads are often used to delimit the cluster boundaries.

### Advantages

Cluster sampling involves selecting population units that are "close" together and does not require all the population units to be listed. Cluster sampling has two advantages:

- it eliminates the need for a complete list of all units in the population; and
- it ensures that selected population units will be closer together, thus enumeration costs for personal interviews will be reduced, and field work will be simplified.

### Disadvantages

In general, cluster sampling is less accurate than SRS (for samples of the same size) because the sample obtained does not cover the population as evenly as in the case of SRS. However it is often preferred because it is more economical.

For example, if we take a simple random sample of 10,000 households across the whole of Australia then we are more likely to cover the population more evenly, but it is more expensive than sampling 50 clusters of 200 households.

### Multi-stage Sampling

Multi-stage sampling involves selecting a sample in at least two stages. At the first stage, large groups or clusters of population units are selected. These clusters are designed to contain more units than are required for a final sample.

At the second stage, units are sampled from the selected clusters to derive the final sample. If more than two stages are used, the process of selecting "sub-clusters" within clusters continues until the final sample is achieved.

The same practical considerations apply to multi-stage sampling as to the cluster sampling.

### Example: A Three-Stage Sample

The following is an example of the stages of selection that may be used in a three-stage household survey.

- Stage 1. Electoral Subdivisions
  Electoral subdivisions (clusters) are sampled from a city or state.

- Stage 2. Blocks
  Blocks of houses are selected from within the electoral subdivisions.

- Stage 3. Houses
  Houses are selected from within the selected blocks.

### Uses of Multi-stage Sampling

Multi-stage sampling is generally used when it is costly or impossible to form a list of all the units in the target population. Typically, a multi-stage sample gives less precise estimates than a simple random sample of the same size. However, a multi-stage sample is often more precise than a simple random sample **of the same cost**, and it is for this reason that the method is employed.

### Advantages and Disadvantages

The advantages and disadvantages of multi-stage sampling are similar to those for cluster sampling.

However, to compensate for the lower accuracy, either the number of clusters selected in the first stage should be relatively large (but this increases the cost of the survey) or the sampling fraction for later stages should be high (i.e. a large percentage of each cluster should be selected).

## SAMPLE SIZE ISSUES AND DETERMINATION

An important aspect of sample design is deciding upon the sample size given the objectives and constraints that exist. Since every survey is different there are no fixed rules for determining sample size. However, factors to be considered include

- the population size and variability within the population;
- resources (time, money and personnel);
- level of accuracy required of the results;
- level of detail required in the results;
- the likely level of non-response;
- the sampling methods used; and
- relative importance of the variables of interest

Once these issues have been addressed, you are in a better position to decide on the size of the sample.

### Variability

The more variable the population is, the larger the sample required to achieve specific levels of accuracy. However, actual population variability is generally not known in advance; information from a previous survey or a pilot test may be used to give an indication of the variability of the population.

When the characteristic being measured is comparatively rare, a larger sample size will be required to ensure that sufficient units having that characteristic are included in the sample.

### Population Size

An aspect that affects the sample size required is the population size. When the population size is small, it needs to be considered carefully in determining the sample size, but when the population size is large it has little effect on the sample size. Gains in precision from increasing the sample size are by no means proportional to population size.

### Resources and Accuracy

As discussed earlier, the estimates are obtained from a sample rather than a census, therefore the estimates are different to the true population value. A measure of the accuracy of the estimate is the standard error. A large sample is more likely to have a smaller standard error or greater accuracy than a small sample.

When planning a survey, you might wish to minimise the size of the standard error to maximise the accuracy of the estimates. This can be done by choosing as large a sample as resources permit. Alternatively, you might specify the size of the standard error to be achieved and choose a sample size designed to achieve that. In some cases it will cost too much to take the sample size required to achieve a certain level of accuracy. Decisions then need to be made on whether to relax the accuracy levels, reduce data requirements, increase the budget or reduce the cost of other areas in the survey process.

### Level of Detail Required

If we divide the population into subgroups (strata) and we are choosing a sample from each of these strata then a sufficient sample size is required in each of the subgroups to ensure reliable estimates at this level. The overall sample size would be equal to the sum of the sample sizes for the subgroups. A good approach is to draw a blank table that shows all characteristics to be cross-classified. The more cells there are in the table, the larger the sample size needed to ensure reliable estimates.

### Likely level of Non-response

Non-response can cause problems for the researcher in two ways. The higher the non-response the larger the standard errors will be for a fixed initial sample size. This can be compensated for by assigning a larger sample size based on an expected response rate, or by using quota sampling.

The second problem with non-respondents is that the characteristics of non-respondents may differ markedly from those of respondents. The survey results will still be biased even with an increase in sample size (ie. increasing the sample size will have no effect on the non-response bias). The lower the response rate, the less representative the final sample will be of the total population, and the bigger the bias of sample estimates. Non-response bias can sometimes be reduced by post-stratification as well as through intensive follow up of non-respondents, particularly in strata with poor response rates.

### Sampling Method

Many surveys involve complex sampling and estimation procedures. An example of this is a multi-stage design. A multi-stage design can often lead to higher variance in resulting estimates than might be achieved by a simple random sample design. If, then, the same degree of precision is desired, it is necessary to inflate the sample size to take into account the fact that simple random sampling is not being used.

### Relative importance of the variables of interest

Generally, surveys are used to collect a range of data on a number of variables of interest. A sample size that will result in sufficiently precise information for one variable may not result in sufficiently precise information for another variable. It is not normally feasible to select a sample that is large enough to cover all variables to the desired level of precision. In practice therefore, the relative importance of the variables of interest are considered, priorities are set and the appropriate sample size determined accordingly.

### Calculation of sample size

When determining an appropriate sample size, we take as a general rule, the more variable a population is, the larger the sample required in order to achieve specific levels of accuracy in survey estimates. However, actual population variability is not known and must be estimated using information from a previous survey or a pilot test. It is worthwhile keeping in mind that the gains in precision of estimates are not directly proportional to increases in sample size (i.e doubling the sample size will not halve the standard error, generally the sample has to be increased by a factor of 4 to halve the SE).

In practice, cost is a major consideration. Many surveys opt to maximise the accuracy of population estimates by choosing as large a sample as resources permit. In complex surveys, where estimates are required for population subgroups, enough units must be sampled from each subgroup to ensure reliable estimates at these levels. To select a sample in this case, you might specify the size of the standard error to be achieved within each subgroup and choose a sample size to produce that level of accuracy. The total sample is then formed by aggregating this sample over the subgroups.

Sample size should also take into account the expected level of non-response from surveyed units.

When the characteristic being measured is comparatively rare, a larger sample size will be required to ensure that sufficient units having that characteristic are included in the sample.

### Sample Size Formulae

If a survey is designed to estimate simple proportions without any cross-classifications in a large population (approximately over 10,000 units), the following formulae can be used to determine the size of the sample:

$$n \geq \frac{p(1-p)}{[SE(p)]^2}$$

where n = sample size,

p = sample proportion,

SE(p) = required standard error of the sample proportion

However, to be able to use this formula, the proportion being estimated needs to be roughly known from supplementary information or a similar study conducted elsewhere. For example, suppose a survey seeks to estimate the proportion of Richmond residents in favour of Sunday night football at the MCG. The standard error (SE) desired is 0.04, while the proportion (p) in favour of the proposal is thought to be about 0.40. The size of the sample would need to be n=150.

If this survey was then completed with a sample size of n=150 and it was a found that the sample proportion (p) in favour of the proposal was 0.8 (not 0.4 as guessed), then the standard error of this sample proportion of 0.8 would be 0.033 not 0.04 as originally planned for.

A proportion of 0.5 gives the highest standard error for a fixed sample size or, requires the highest sample size for a fixed standard error, hence p=0.5 is the worst case scenario. It is for this reason that an estimate of p=0.5 is often used when calculating sample sizes when there is no information on the proportion to be estimated.

**Example: Gains From Sampling**

Suppose we wish to take a sample from a population. We have a preliminary estimate of the proportion of the population having the characteristic we are interested in measuring (50%). The level of accuracy we require from our survey is an RSE of 5%. Using the formulae for the sample size in a finite population.

For various population sizes, the sample size that we would need is:

| Population | Sample Size | Sample Fraction (%) |
|---|---|---|
| 50 | 44 | 88.000 |
| 100 | 80 | 80.000 |
| 500 | 222 | 44.400 |
| 1,000 | 286 | 28.600 |
| 5,000 | 370 | 7.400 |
| 10,000 | 385 | 3.850 |
| 100,000 | 398 | 0.400 |
| 1,000,000 | 400 | 0.040 |
| 10,000,000 | 400 | 0.004 |

The gains from employing sampling are greatest when working with large populations.

This page last updated 1 March 2023

## QUESTIONNAIRE DESIGN

### Introduction

A poorly designed questionnaire can be the biggest source of non-sampling error (either directly or indirectly). The questionnaire can influence the response rate achieved in the survey, the quality of responses obtained and consequently the conclusions drawn from the survey results.

Questionnaire design should be started by considering the objectives of the survey and the required output, and then devising a list of questions to accurately obtain this information. Careful consideration should be given to a number of factors including the types of questions to be asked, the questionnaire wording, the structure and design of the questionnaire and testing the questionnaire to ensure that quality data is collected.

This section discusses the developmental phase of questionnaire design, question wording, different types of questions and the structure and design of the questionnaire. Testing the draft questionnaire is covered in Survey Testing.

### Questionnaire Development

Before a questionnaire can be designed, time must first be spent clarifying the objectives of the survey with the client, researching the subject matter and methodology of the collection and conducting testing to obtain an insight into the characteristics of different groups of the target population.

This development phase needs to be undertaken to ensure that the data items in the final questionnaire are appropriate and can be collected accurately. It is also important to consider early on, the coding of questionnaire responses and the type of processing system to be used to process the questionnaire.

### Defining Collection Objectives with the Client

The Set-up State of a Survey outlined the importance of clearly establishing the research objectives for the collection of any information. Similarly, in the process of questionnaire development, the reasons for collecting the information need to be clearly articulated. Defining the collection objectives with the client includes:

- clarifying the objectives of the survey;
- justifying the collection in terms of the benefits in the collection and its public good;
- determining the scope of the survey, ie who is to be surveyed;
- determining the desired output, ie specify the tables according to the objectives and the precision required;
- preparing a list of content, ie information needed according to the objectives. A number of questions may be necessary to obtain the desired information;
- defining the content;
- justifying the content, ie is all the information collected necessary; and
- establishing priorities for each data item. This is important in ensuring that the most important data items are collected. It also makes it easier to discard less important questions if the survey budget is reduced.

A similar process of defining and justifying the content in consultation with users should be conducted when adding new or redeveloped questions to an existing collection.

### Researching the Topic

As stated in The Set-up Stage of a Survey, it is very important that the researcher who is going to design the questionnaire undertakes background research into the topic under consideration. In terms of questionnaire design, the research should aim to:

- clearly define the concepts and definitions to be used in the survey. The use of standard definitions and classifications will enhance data comparability; and
- accurately identify the target population characteristics. This needs to be done so that questions can be pitched at an appropriate level.

It is also worthwhile looking at past collections of the same topic to learn from past experience. Such collections can prove to be a useful basis for the current collection and also help to avoid possible mistakes. It is important to look at the concepts, definitions and question wording of past collections if a time series is to be created, since changes in these will result in changes in responses. For example, a question in the 1981 Population Census, 'Do you have a mortgage?' received very different responses from those that were obtained in the 1986 Census when the question was changed to 'Have you paid off your house?' Many people who had second mortgages included this in the 1981 Census but not in the 1986 Census.

Past collections of similar topics may also be worth looking at. For example, if you are collecting information on drug use it may be appropriate to follow a similar methodology to that used in surveys of alcohol use.

In examining past collections the method of collection used should also be considered. Some topics will be more or less appropriate for particular method.

**General Considerations**

The main functions of questionnaires are to extract information from respondents, record and process it. To achieve this effectively, a questionnaire should:

- provide clear instructions (for respondents, interviewers and processors);
- clearly and concisely define what is to be collected and recorded;
- maintain respondents' cooperation and involvement;
- enable respondents to complete it accurately and within a reasonable time;
- use a language that is understood by the respondents;
- avoid bias in question wording;
- make the job of the respondent and/or interviewer easy;
- appear uncluttered on the form;
- provide suitable space for responses; and
- be easily processed by both people and machines.
- be in a suitable form for keeping as a hard copy record.

Additionally, it is necessary to realise that a questionnaire often cannot provide all the information that a client would like to obtain. If a questionnaire becomes too long or confusing, respondents may be unwilling to complete it, or they may make mistakes.

## TYPES OF QUESTIONS

Questions can be classified into types based on their answer formats and also the type of data sought.

### Answer Formats

Questions can generally be classified as one of two types - open or closed - depending on the amount of freedom allowed in answering the question. When choosing the type of questions, consideration should be given to factors such as the kind of information which is sought, ease of processing, and the availability of the resources of time, money, and personnel.

**Open Questions**
Open questions allow the respondents to answer the question in their own words. An example is 'What is your occupation?' The advantages of these type of questions is that they allow many possible answers and they can collect exact values from a wide range of possible values. However, they are more demanding than closed questions, both to answer and process. Open questions are often used in pilot tests to determine the range of likely responses.

**Closed Questions**
Closed questions provide respondents with a range of the most likely answers to choose from. These questions are appropriate when the researcher can anticipate most of the responses and when exact values are not needed. However, they require more effort than open questions in the development and testing stages. Processing time of closed responses is much less than that of open ended responses. Some examples of types of closed questions are given below:

- Limited Choice

Limited choice questions are those which require a respondent to choose one of two mutually exclusive answers. For example, yes/no.

- Multiple Choice

These are questions where the respondent is required to choose from a number of responses provided.

- Checklist

  Checklist questions allow a respondent to choose more than one of the responses provided.
- Partially Closed

  These questions provide a set of responses where the last alternative is 'Other, please specify'. Partially closed questions are useful when it is difficult or impractical to list all possible choices.

### Choosing Between Questions Types

In choosing between these two alternatives (open or closed questions), consideration should be given to factors such as the data requirement, the kind of information sought, the level of accuracy needed, ease of processing and the availability of coding resources, the position of the questions on the form and the sensitivity of the question. In general, closed questions are better for both the interviewer and the coder.

Once the questions have been chosen, they should be tested and retested until the best choice has been made. This is covered in more detail in Survey Testing.

### Type of Data Sought

There are six main types of information or data that can be obtained from questions. They are discussed below.

**Factual Questions**
In these questions, factual information is required of the respondent rather than an opinion. Respondents could be asked about possession of items (eg. 'Do you have a driver's licence?') or characteristics of the business (e.g. 'How many employees does this business have?').

**Opinion or Motivational Questions**
These questions seek opinions ('Are you in favour of a capital gains tax?') rather than facts. There are many problems associated with opinion questions. For instance:

- a person's attitude to a subject may not be fully developed or they may not have given it much thought;
- opinion questions are very sensitive to changes in wording; and
- it is impossible to check the validity of responses to opinion questions.

**Behavioural Questions**
These questions require information about the activity of the respondent or business (e.g. 'How many times did you go to the theatre in the last 12 months?'). Behavioural questions need to be used with care because they often require difficult recall tasks on the part of the respondent. They should be restricted to topics respondents will remember easily or are likely to have records for, and cover a reasonable and specific time frame.

**Hypothetical Questions**
The "What would you do if ... ?" type of question. The problems with hypothetical questions are similar to opinion questions. You can never be certain how valid any answer to a hypothetical question is likely to be. They should only be used to refer to a hypothetical *occurrence* of a type of situation a respondent will be familiar with or have structure in place for. An example of this is for testing questions e.g. 'If an employee went on holidays at the end of January and was paid in advance for all of February, would you include them in the number of employees reported for the pay period ending on or before 21 February?'

**Classification or Demographic Questions**
These are used to distinguish the main groups of respondents in a survey for later analysis (eg. age, sex, industry). They are usually left towards the end of a questionnaire unless they are necessary for *filter questions* (questions which direct respondents to skip questions that do not apply to them).

**Knowledge Questions**
These questions test the respondent's knowledge about current issues etc. For example, 'Who is the Prime Minister?', 'Are you aware of these industry support groups?'

## QUESTION WORDING AND RESPONSE CATEGORIES

There are a number of factors to consider when designing questions to ensure that appropriate answers are obtained. Several aspects of question design can introduce error, namely:

### Language

Questions which employ complex or technical language or jargon can confuse or irritate respondents. In the case of interviewer based surveys, respondents who do not understand the question may be unwilling to appear ignorant by asking the interviewer to explain the questions. The respondent may then either refuse to answer or give an inaccurate response.

Technical language or jargon should only be used in cases where it is part of the normal language of the survey's target population. An example of this case would be a survey of information technology specialists: the survey would need to use language that is 'jargon' to the survey designer, but appropriate for the respondent.

A general principle to keep in mind is that the wording of questionnaire items should be specific, definitive, consistent, brief, simple and self-explanatory.

### Ambiguity

If ambiguous words or phrases are included in a question, the meaning may be interpreted differently by different people. This will introduce errors in the data since different respondents will be virtually answering different questions.

For example, consider the question 'Has your standard of living decreased substantially because of a sharp increase in your monthly mortgage repayments?' A 'No' answer could mean any one of a number of things - for instance: 'No my standard of living has not dropped because of increased repayments' or 'No, my repayments have not increased'.

A question may also seem straightforward, but allow for a variety of different kinds of answers. It is important to include the measurement unit you require wherever one applies, e.g. dollars, days, litres.

### Double-Barrelled Questions (Multiple Concepts in one Question)

These are apparently single questions which actually incorporate two different questions. For example: ' Do you intend to leave work and return to full-time study this year?' A person may be intending to leave work, but not to return to study, and vice versa. When different parts of the question have different answers, or parts of the question are not relevant, respondents may be unsure how to answer. When attempting to interpret answers to such questions, it can be unclear to which part of the question the answer corresponds.

### Leading Questions

Error will be introduced if questions lead respondents towards a particular response. For example, the question 'How many days did you work last week?', if asked without first determining whether respondents did in fact work in the previous week, is a leading question. It implies that the person would have or should have been at work. Respondents may answer incorrectly to avoid telling the interviewer that they were not working.

### Unbalanced Questions

Another form of leading questions are unbalanced questions. For example, 'Are you in favour of gun control?' provides only one alternative to consider. The question should be reworded to something like 'Do you favour gun control, or are you against gun control?', which gives respondents more than one alternative. The answer options of a question can also be unbalanced. For example, a respondent could be asked in a neutral way "Please rate your overall health" but required to select from the answers "Poor", "Good" and "Excellent".

### Recall/Memory Error

A significant degree of error can be introduced in questions which require respondents to recall events, expenditure etc, particularly if details are being sought for a long period. The quality of the data collected from recall questions is influenced by the importance of the event to the respondent and the length of time since the event took place. Respondents also tend to remember what should have been done rather than what was done.

Subjects which are of greater importance or interest to respondents, or events which happen infrequently, will be remembered over longer periods and more accurately. Where possible (eg. with financial information), questions should be framed so that respondents can refer to their own records which would enhance accurate reporting. Minimising the recall period also helps to reduce memory bias.

A specific type of memory error is telescoping. This occurs if respondents report events as occurring either earlier or later than they actually occur, incorrectly bringing events into the reference period. This effect is alleviated somewhat by being very specific about when the reference period begins and ends, for example using "the week ending Saturday 1st September" rather than "last week".

### Intrusive (Sensitive) Questions

Questions on topics which respondents may see as embarrassing or highly sensitive can produce inaccurate answers. Respondents may refuse to provide information on personal issues such as health or income details. If respondents are required to answer questions with information that might seem socially undesirable, they may provide the interviewer with responses they believe are more 'acceptable'. In these cases it is often better to provide the respondent with a self-administered questionnaire that the interviewer doesn't see.

Business survey respondents can also find some topics sensitive, such as IT security breaches or donations to charity, as well as not wanting to reveal commercial-in-confidence information about their business. Business surveys add a new dimension to collecting sensitive data as it is often necessary for different respondents, sometimes from different areas, to complete parts of the form and approve its content. Some respondents might not want the others to see particular answers.

The negative effect of sensitive questions may be aggravated if they are placed at the beginning of the questionnaire and can therefore contribute to non-response if respondents are unwilling to continue with the remaining questions. If a sensitive question is further into a form the respondent is more committed to completion, and if they do refuse to continue, the partial response is more useful. Ways of overcoming difficulties associated with sensitive questions may include reassuring respondents that the information they provide is confidential, and not requiring respondents to write their name anywhere on the survey form.

### Attitude Strength

These questions seek to locate a respondent's opinion on a rating scale with a limited number of points. For example, a five point scale measures strong and weak attitudes, ie respondents may be asked whether they strongly agree/agree /neither agree nor disagree/disagree/strongly disagree with a given statement. Whereas a three point scale would only measure whether they agree, disagree, or neither, but not the strength. Using many scale points will generally not provide meaningful differences in attitude strength.

Care needs to be taken with the use of attitudinal scales because some respondents may have difficulty interpreting the scale. Additionally, such scales are interpreted subjectively and this interpretation can differ between respondents.

### Acquiescence

This situation arises when respondents have a general tendency to agree rather than disagree with anything. It occurs when respondents are asked whether they agree or disagree with a statement, especially when the supplied statements are presented as plausible generalities. It can also appear for questions requiring a yes or no response.

This tendency can be due to a combination of factors, such as the personality and education level of respondent, as well as conditions of the interview or design of a self-completed questionnaire. Respondents will often agree when the question is ambiguous or otherwise difficult to answer. The effect may be exaggerated when the respondent is fatigued or has to answer a long string of questions with the same response categories. A related effect is satisficing, where respondents select the first reasonable answer rather than make the effort to find or remember the best answer.

### Adequate Response Categories

It is important to make sure that there are adequate response categories and that they incorporate every possible response. For example:

Age 15-19 21-25

This provides a problem for those respondents whose age is 20.

Another problem that could arise is overlapping response categories. Ranges should always be mutually exclusive. For example:

Age 15-20 20-25

This provides a problem for respondents whose age is 20 since they could respond in either or both categories.

Response categories also need to be worded carefully, as respondents will use them to clarify or extend the meaning of the question. For example, if a question used a frequency scale with five points "Never", "Rarely", "Average", "Often", and "Frequently" a respondent may incorrectly assume the scale represents the population distribution. If they consider themselves to be normal or extreme compared to the population on the activity of interest their answers will differ regardless of the actual frequency they engage in the activity.

### Number of Response Options

The number of response categories can influence the quality of the data as both too few and too many categories can cause errors. Too many can cause respondent fatigue and inattention, resulting in ill-considered answers. If there are too few categories respondents may have difficulty finding one which accurately describes their situation.

### Don't Know Category

The decision about whether to include or exclude a Don't Know option depends to a large extent on the subject matter. The remaining responses are usually evenly distributed on the negative and positive sides of a scale if this category is excluded, although it depends to a large extent on the nature of the question. Excluding this option may not be a good idea, as respondents may be forced to give an answer when, for example, they really do not know what their attitude is to a particular subject, or they do not know the answer to a factual question that has been asked. When respondents are forced to develop an attitude on the spot this attitude will be highly unreliable.

### Tone

A change in wording can result in a change in responses. For example, different response may be obtained through using the following two questions:

'Do you think that gun ownership should be forbidden?' or

'Do you think that gun ownership should not be allowed?'

Minor changes in wording can also have a significant affect on responses. One should therefore be careful when looking at alternative wordings. The use of negative words like "not" should be avoided in questions as they are easily missed by respondents. In addition, using "not" in a scale such as "Satisfied", "Neither" and "Not satisfied" doesn't provide a true opposite. "Dissatisfied" would be a better alternative, however "Unsatisfied" could also be used and would mean something slightly different to respondents.

## STRUCTURE AND LAYOUT OF THE QUESTIONNAIRE

Not only does the wording of questions require attention to detail, but also the 'look' of the questionnaire. Poorly designed questionnaires (eg. hard to read text) not only serve as a disincentive to respondents completing the questionnaire, but can also result in respondents making errors. Some of the important elements of questionnaire structure and layout are outlined below.

### Sequencing

The questions on a form should follow a sequence that is logical to the respondents. Regardless of the method used to administer the questionnaire, the sequence should flow smoothly from one question to the next. A smooth progression through the questions is particularly important if the questionnaire is answered in difficult circumstances (e.g. a mother trying to fill in a questionnaire while her children are seeking her attention). It is a good idea to start the questionnaire with pleasant and easy questions to promote interest in the survey, and to give respondents confidence in their ability to answer the remaining questions. In particular, the opening questions should establish that the respondent is a member of the survey population.

The remaining questions should be logically structured so that the interviewer or respondent does not need to alternate between pages of the questionnaire. For example, any explanatory notes should be presented as part of the question they refer to, not on a separate page.

Questions which may be sensitive to respondents should generally not be placed at the beginning of a questionnaire. Rather, they should be placed in a section of the form where they are most meaningful to the context of other questions. In this way the format of the questionnaire can act as a buffer to help the respondent feel more comfortable with sensitive questions after establishing rapport.

In self-enumeration questionnaires, to ensure that respondents answer only those parts of the questionnaire that are relevant, filter questions may be used to direct respondents to skip the questions that do not apply to them. Filter questions are also used in interviewer based surveys to direct interviewers to follow a series of questions according to answers given by respondents. Filter questions need to be used with care as respondents (and interviewers) need to have sufficient information about the skip condition to judge whether the respondent should skip. Filters should also generally be avoided for sensitive topics as respondents will tend to give the answer that avoids answering the sensitive questions.

If the instructions are not clear and straightforward, interviewers or respondents can follow an incorrect sequence or miss questions. In general, only one or two conditions should be placed in each sequence guide. Computer-assisted interviewing and electronic self-completion forms can make complex sequencing much easier.

Filter questions also identify sub-populations. For example:

Q7 'Were you born overseas?'
If 'Yes' go to Q8,
if 'No' go to Q12.

### Order of Questions
The order in which the questions appear may influence the responses given to particular questions. Responses given to earlier questions can influence responses to later questions. For example, if a question asks participants whether they believe trade unions are disruptive in the community, and then a later question asks about problems in Australian industry, the negative influence of trade unions on industry may receive much more attention than would otherwise have been the case.

### Order of Response Options
The actual order of response options can also introduce bias. The options presented first may be selected because they make an initial impact on respondents, or because respondents lose concentration and do not hear or read the remaining options. Equally, the last options may be chosen because they are more easily recalled, particularly if respondents are given a long list of options. Thus, the order of response options has a greater effect on data quality when a question includes a large number of response options.

If possible, options should be presented in a meaningful order. If some options are more socially desirable than others these should go last to reduce bias. For example, an education question should present the qualifications in order from lowest to highest. For some self-completed lists, alphabetical order is the most appropriate to help the respondent find which option they want, for example if respondents have to select which crops they produce.

### Response Options and Respondent Difficulties
When the survey is interviewer based, the response options can be presented either verbally or on a prompt card. A prompt card is a list of possible responses to a question which are shown by the interviewer to assist the respondents. This helps to decrease error resulting from respondents being unable to remember all the options read out to them. However, respondents with poor eyesight, migrants with limited English, or adults with literacy problems will experience difficulties in answering accurately.

### Length
The length of a questionnaire can be described in different ways. Survey designers tend to worry about number of pages, whereas the number of questions (especially mandatory ones) and the time taken to complete are usually more important. How long is too long varies across mode of data collection, the ease and interest of the topic and the design of the questionnaire. Towards the end of a long questionnaire, respondents may give less thought to their answers and concentrate less on the instructions and questions, thereby decreasing the accuracy of the information they provide. If a respondent is told an interview will last several hours, or they receive a questionnaire that is many pages thick, that can lead the respondent to refuse to participate at all.

### Questionnaire Layout
Respondents or interviewers using a questionnaire with poor layout can miss questions, follow an incorrect sequence or enter responses in the wrong response box, which will result in missing or incorrect data. Poor layout can also contribute to errors at the processing stage. For example, if response boxes are not aligned some answers may be missed completely during the data entry.

Particularly for those questionnaires which are completed by respondents, the design may contribute to errors as a result of:

- poor legibility (e.g. unclear printing or very small text);
- violating the normal reading path of the respondent (English readers expect to read from top left to bottom right and e.g. big headings and bright pictures in the middle of pages disrupt this);

- instructions which can be easily overlooked (e.g. those which are not clearly differentiated from the questions or are not placed near the relevant part of the question); and
- inadequate space for answers.

A questionnaire should be built up question by question. That is, it is inappropriate to start with the page and then try and make the questions fit.

**Physical Design**

The questionnaire should be physically set out so as to minimise the time needed to interview, respond and process the results. Specifically, consideration should be given to the form's construction, graphics, and layout. Poor layout leads to mistakes in understanding questions and recording the responses. In general, the questionnaire:

- needs to be understood by respondents, interviewers and processors. This is done by providing clear instructions. For respondents this means adequate layout of questions and response categories with appropriate sequencing. Interviewers require prompts to be clearly understood and response codes need to be adequate for processors;
- should have a good appearance as it might affect the response, i.e. the questionnaire should be well designed and presented and therefore easy to answer;
- should clearly identify the date, title and the organisation;
- should clearly outline the purpose of the survey;
- should assure respondents about the confidentiality of the information they are providing;
- should provide a contact number so that respondents can obtain help if they require it and a due date; and
- should have pages that are numbered consecutively, with a simple numbering system.

Specific principles of good questionnaire design are outlined below:

- left align text where possible;
- typeface: avoid ornate and decorative typefaces. Serif type fonts (eg. Times) are easier to read for questions (on paper) than sans serif types (eg. Helvetica);
- upper case text is difficult to read, and should be avoided where possible;
- allow enough space for answers;
- tick boxes are a popular way of obtaining responses. Where possible, they should be vertically aligned rather than horizontally aligned; and
- lines are sometimes useful in order to divide columns, sections and questions. These lines should be as fine as possible and they should only be used where necessary;
- keep the amount of ink on the form to the minimum necessary for the form to work properly. Cluttered forms contribute to respondent fatigue and errors, thus leading to a decline in data quality;
- black text is easiest to read;
- the background colour of the questionnaire form should not be too strong or bright. Light pastel colours are usually suitable. Background colour should contrast enough with the text to allow for ease of reading;
- avoid colour combinations such as red and blue or green and orange. Black on pale yellow is a good combination, but it is difficult to get the right yellow.

**Testing Questionnaires**

Once the understanding of concepts and definitions has been investigated through focus groups, a rough questionnaire can be produced and tested informally on a small group of people, perhaps one's colleagues at work. Such testing is not intended to obtain representative results, but aims to find out the major flaws with the questions, for example awkward wording. This testing is designed to take the 'rough edges' off a questionnaire. It is a good idea to use open questions to work out the likely responses. The questions can be restructured and developed into a draft questionnaire which can be used in rounds of informal pretesting and later pilot testing. For further details Survey Testing.

**Conclusion**

Questionnaire design begins by clarifying the objectives of the survey, determining the data which is to be produced by the survey and devising a list of questions to obtain this data. Careful consideration should be given to a number of factors, including the type of questions to be used, the logical sequence and wording of questions, and the physical design of the form. It is important to test each of these aspects of questionnaire design with a group of respondents before finalising the questionnaire. If necessary, the form can then be modified and retested until respondents can complete it accurately and quickly with a minimum of errors.

This page last updated 1 March 2023

## SURVEY TESTING

### Introduction

Survey testing is an important part of the survey development process and provides guidance on the :

- adequacy of the sampling frame
- variability of the target population with regard to the survey subject
- expected non-response rate and the effectiveness of measures aimed at reducing non-response
- suitability of the data collection method, including testing of various methods to determine the most suitable
- adequacy of the questionnaire, including testing of alternative versions to determine the most effective
- effectiveness of interviewer training and the adequacy of survey instructions
- answer categories to be used for pre-coded questions
- likely cost and duration of parts of the survey and the survey as a whole
- organisation of the survey.

In some cases, data collected in the tests may be useful preliminary indicators of the survey results; estimating sampling error, sample sizes and population variability; and estimating likely response rates. These preliminary results will involve a smaller sample and thus produce higher standard errors, which need to be taken into consideration.

## TYPES OF TESTING

There are two main types of testing that can be used to evaluate a survey: Qualitative testing and Quantitative testing. Within qualitative testing are techniques such as skirmishing, focus groups, pretesting and observational studies. Pilot testing and dress rehearsals are quantitative tests. Each type of testing is used at a different stage of the survey's development and aims to test different aspects of the survey.

### Skirmishing

Skirmishing is the process of informally testing questionnaire design with groups of respondents. The questionnaire is basically unstructured and is tested with a group of people who can provide feedback on issues such as the questions' frame of reference, the level of knowledge needed to answer the questions, the range of likely answers to questions and how answers are formulated by respondents.

### Focus Groups

Focus groups are small groups of people who represent or who share similar characteristics with the target population. Focus groups are used to test concepts or questions relating to a small segment of the population, who may not be adequately represented in a larger pilot test.

Focus groups are usually small as large groups can be difficult to control. They need to be led by an experienced facilitator whose job is to control the group discussion and ensure that all participants have the opportunity to put their views forward. In conducting focus groups you should address the following issues:

- The relevant subject matter experiences of the group,
- The characteristics of the target population,
- How well do the participants understand the concepts and definitions used,
- The abilities of the group (e.g. can they add numbers if asked), and
- What is the general group reaction to the form wording and layout.

Focus groups can help us to refine our image of respondents, both in general terms and in relation to significant sub-groups of our survey population.

Informal discussions are held with the focus groups to gain an insight into the attitudes, concerns and responses of the group, related to the questionnaire. When introducing a new questionnaire, it is quite useful to gather a number of likely respondents together and discuss the issues involved. Out of these discussions can come such things as the availability of the data and details of the people best suited to supplying it.

The volume of quite specific information obtained from focus groups means that, depending on the nature of the concepts or questions being investigated, only one or two focus groups are required.

Due to the way focus groups examine very specific issues, they should not be used as the sole method of testing a new or revised questionnaire. Another danger with focus groups results from a phenomena called 'group speak', where problems can be built up to exaggerated proportions as members of the group 'feed off' each other during discussions.

### Pretesting (also known as Skirmishing)

Pretesting is the process of informally testing questionnaire design with potential respondents. The questionnaire is basically unstructured and is tested with a group of people who can provide feedback on issues such as the questions' frame of reference, the concepts, the level of knowledge needed to answer the questions, the range of likely answers to questions and how answers are formulated by respondents. Pretesting is also used to detect obvious flaws or awkward wording of questionnaires as well as testing alternative designs. At this stage we may use open-ended response categories to work-out likely responses. The questionnaire should be redrafted after pretesting.

### Observational Studies

Observational studies involve getting respondents to complete the draft questionnaire in the presence of an observer. Whilst completing the form, respondents explain their understanding of the questions and the methods required in providing the information. Much can be gained from such studies including identifying problem questions through observations, questions asked by the respondents, or the time taken to complete particular questions. Data availability and the most appropriate person to supply the information can also be obtained through observational studies. Respondents should be made aware that it is the form that is being tested and not the respondent. The respondent should not be given assistance in completing the form.

### Pilot Testing

Pilot testing involves formally testing a questionnaire or survey with a small representative sample of respondents. Semi-closed questions can be used to gather a range of likely responses which are later used to develop a more highly structured questionnaire with closed questions. Pilot testing is used to identify any problems associated with the form, such as questionnaire format, length, or question wording, and allows comparison of alternative versions of a questionnaire. Pilot testing can also be used to assess the adequacy of instructions to interviewers and ascertaining interview times. For a new survey, several pilot tests may be necessary before the survey methodology is considered satisfactory.

### Dress Rehearsals

A dress rehearsal is a final trial run of the survey where the chosen sampling methodology is used to select a small sample from the target population. Dress rehearsals are used to detect any problems that may arise in the survey design and/or processing system. Dress rehearsals also provide an opportunity to obtain data on survey costs and estimate population variances. Administrative costs, including average interview times as well as cash outlays, are obviously useful in deciding staffing and budgetary arrangements. Estimates of population variances appropriate to the sample design (or designs) under consideration may be needed to predict the sampling errors of any proposed sample plans.

If appropriate information is available from previous surveys, it may be possible to estimate population variances and costs from this information rather than from the results of a dress rehearsal.

## CONDUCTING SURVEY TESTS

### Composition of Survey Tests

Once testing procedures have been decided upon, the next step is to determine how these tests will be run. The samples chosen for pilot testing and dress rehearsals need to be as representative of the target population as possible. This maximises the validity of the test results and ensures that consequent modifications to the survey are appropriate. The sample may be chosen using judgement sampling to obtain a fairly representative sampling method or using probability sampling methods. Observation studies and focus group testing do not need representative samples for them to be carried out.

### Comparison of Techniques

Comparisons between pilot tests and dress rehearsals can be used as a basis for choosing between alternative procedures for part of a survey. One approach to such testing could be to allocate two equal size samples to interviewers and each interviewer uses both of the alternative procedures. The two techniques can then be compared without the interfering bias of differing interviewer performance.

### Frequency and Size of Tests

The number of times survey testing should be conducted and the sample sizes used are determined by the complexity of the survey and the availability of funds. For example, a simple, small survey may need only 50 respondents for a dress rehearsal, whereas a larger, more complex survey may need 200 or more respondents.

## ANALYSIS OF TEST RESULTS

The results of survey testing can bring to light a number of problems with the questionnaire and survey design. Some of these problems may be identifiable during pre-testing; others may be identifiable only when the questionnaire is administered to a sample of the target population during a pilot test or dress rehearsal.

### Problems with Questionnaires

Lengthy questionnaires will often intimidate respondents into rushing through questions or refusing to continue. This can lead to a high percentage of non-response and a general misunderstanding of questions. Poor layout and explanation of questions can also lead to non-response for questions or multiple answers being given when only one should be chosen. Other problems that may be identified include a large number of 'Other' responses to a particular question, lack of variation in the responses to a question and general misunderstanding of a question.

### Solutions to Questionnaire Problems

The solution to these problems may lie in the sequencing of questions, the instructions accompanying questions, the wording of questions and the answer categories provided. Changing the answer categories can produce more variation in the responses given to a question. We can do this by increasing the list of answer categories, changing the order of the list or even changing the emphasis of the question and/or the answer categories.

### Survey Design Problems

In addition to problems with questionnaire design, survey testing may also uncover problems with the sampling methodology and administrative calculations. These may include deficiencies in the sampling frame and/or sample size chosen, sources of sampling and non-sampling error, problems with the processing system and both budgetary and time constraints. Once identified, these problems can be addressed and either eliminated or minimised before the full survey is conducted. When used in this way, pilot testing is an invaluable tool for maximising, within resource constraints, the quality of results obtained from the final survey.

### Preliminary Survey Results

The results of pilot testing can also be used to provide a 'preview' of the results of the full survey. The data gained from the test can be analysed in the same way as the final survey, incorporating tables, statistical analyses (e.g. correlations, scales, etc.) and discussion of the findings. The data will be subject to higher standard errors although it may still be useful.

This page last updated 1 March 2023

## DATA PROCESSING

### Introduction

Data processing involves translating the answers on a questionnaire into a form that can be manipulated to produce statistics. In general, this involves coding, editing, data entry, and monitoring the whole data processing procedure. The main aim of checking the various stages of data processing is to produce a file of data that is as error free as possible. Adopting a methodical and consistent approach to each processing task is important to ensure that the processing is completed satisfactorily and on schedule. The following discussion will give a brief introduction to the main stages of data processing.

### Stages in Data Processing

Functions involved in data processing include: despatch and collection control; data capture and coding; and editing.

### Despatch and Collection Control

A despatch and collection control (DACC) system can be employed to organise the distribution and receipt of the survey forms. It is also used to employ mark-in functions, employ reminder action, and, importantly, generate management information reports on the collection's progress.

All of the above tasks could be performed manually but are more likely to be automated using some kind of generalised system if the survey is large. The functionality required in the DACC phase can vary significantly, depending on the collection technique, i.e. whether a mail survey, personal interview, telephone survey or a combination of techniques are used.

DACC segments include population file set-up and maintenance, label production, hardcopy/electronic collection control registers, mark-in, reminder control, intensive follow-up control and management information. Management information covers response rate progress, returns to sender, timing and effects of each phase such as reminder action.

### Data Capture and Coding

The questionnaire can be used as a working document for the transfer of data on to a computer file, therefore removing the need for a separate data input form. This also removes a stage that could produce transcription errors. Consequently, it is important to design the questionnaire to facilitate data entry.

### Coding

Unless all the questions on a questionnaire are "closed" questions, some degree of coding is required before the survey data can be sent for punching. The appropriate codes should be devised before the questionnaires are processed, and are usually based on the results of pilot tests.

Most surveys are too large and complex to be analysed by checking questionnaires and counting responses. Surveys usually require a system by which the responses can be transferred onto a computer file for analysis. This system would involve translating responses from "closed" questions into numerical codes.

Coding consists of labelling the responses to questions in a unique and abbreviated (using numerical codes) way in order to facilitate data entry and manipulation. Codes should be formulated to be simple and easy, for example, if Question 1 has four possible types of responses then those four responses could be given the codes 1, 2, 3, and 4. The advantage of coding is the simplistic storage of data as a few-digit code, compared to lengthy alphabetical descriptions which almost certainly will not be easy to categorise.

Coding is a relatively expensive task in terms of resource effort. However, improvements are always being sought by developing automated techniques to cover this task.

The coding frame for most questions can be devised before the main interviewing begins or forms are despatched. That is, the likely responses are obvious from previous similar surveys or through pilot testing, allowing those responses and relevant codes to be printed on the questionnaire. An "Other" answer code is often added to the end of a coding frame with space for respondents or interviewers to write the answer.

A major function of the ABS is the development of statistical standards, classifications and frameworks. These classifications cover a wide range of subjects, including the:

- ASGC Australian Standard Geographical Classification (Cat. No. 1216.0);

- ASCO Australian Standard Classification of Occupations (Cat. No. 1222.0);
- ANZSIC Australian & New Zealand Standard Industrial Classification (Cat. No. 1292.0); and
- ABS Standards for Statistics on the Family (Cat. No. 1286.0).

ABS classifications should always be considered when designing questions for a questionnaire because they are easily coded and allow for comparison with ABS data.

### Data Capture

Data capture methods are very expensive in terms of both staff and time. Some of the advances in recent times that reduce the number of errors introduced by data capture are given below.

In a mail-based survey errors may occur when clerical staff enter the data using a computer terminal. These errors can often be reduced by improving the design of the data entry system. This is increasingly possible with the introduction of computer aided data entry (CADE) systems which can be tailored to the form being entered. A good CADE system would present screen images of pages from the paper form, making it easy for the clerk to enter the respondent's values at the appropriate points.

Another approach to minimise data entry is by using machine-readable returns. Optical mark reading (OMR) can be used to read form identifiers, and in some cases to read responses consisting of marks. Optical character recognition (OCR) allows the computer to read hand written forms directly.

In interviewer-based surveys errors may occur as the interviewer fills out the form and also later when the data is entered into the computer. If the double transfer of data can be avoided the opportunity for errors can be reduced. Approaches to this are computer assisted telephone interviewing (CATI) and computer assisted personal interviewing (CAPI) which involve the interviewer entering data directly into a computer terminal rather than onto a form.

### Factors to Consider

When changing to a new data capture system, there are several issues that need to be considered. These include the accuracy of results, the time in which results can be output, the staffing required for the new system and the costs of the changeover. In particular, some of the considerations might be:

- in on-going surveys, a change in system can cause a drift in the results, not allowing comparability between different survey periods;
- a reliance on a small number of data capture machines can cause timeliness problems when breakdowns occur;
- there may be a decrease in the need for data punchers as far as staff levels are concerned if the data capture machines process forms more quickly;
- when changing systems, there are large up-front costs in introducing hardware, but improvements in the capture system should lower staff costs.

## Editing

### What is editing?

Editing is the process of correcting faulty data, in order to allow the production of reliable statistics.

Data editing does not exist in isolation from the rest of the collection processing cycle and the nature and extent of any editing and error treatment will be determined by the aims of the collection. In many cases it will not be necessary to pay attention to every error.

Errors in the data may have come from respondents or have been introduced during data entry or data processing. Editing aims to correct a number of non-sampling errors, which are those errors that may occur in both censuses and sample surveys; for example, non-sampling errors include those errors introduced by misunderstanding questions or instructions, interviewer bias, miscoding, non-availability of data, incorrect transcription, non-response and non-contact. But editing will not reveal all non-sampling errors - for example, while an editing system could be designed to detect transcription errors, missing values and inconsistent responses, other problems such as interviewer bias may easily escape detection.

Editing should aim:

- to ensure that outputs from the collection are mutually consistent, for example, a component should not exceed an aggregate value; two different methods of deriving the same value should give the same answer;
- to detect major errors, which could have a significant effect on the outputs;

- to find any unusual outputs and their causes.

**The required level of editing**
The function of editing is to help achieve the aims of a collection so, before edits are created or modified, it is important to know these aims - since these have a major say in the nature of the editing system created for the given collection. We need to know about features such as:

- the outputs from the collection
- the level at which outputs are required
- their required accuracy
- how soon after the reference period the outputs are needed
- the users and uses of the collection. A collection may be simple (with limited data collected) and designed to meet the requirements of only one type of user (e.g., Survey of new Motor Vehicle Registration and Retail Trade) or it may collect more complex data and aim to meet the needs of many different types of users (e.g. Agricultural Finance Survey, Household Expenditure Survey, etc.). If there are many types of users there is a likelihood of conflicting requirements amongst the users, which can lead to stresses on the collection.
- the reliability of each item (eg. is the definition easily understood or is the item sensitive?)

While the goal of editing is to produce data that represent as closely as possible the activity being measured, there are usually a number of constraints (such as the time and number of people available for data editing) within which editing is conducted. These constraints will also influence the design of the editing system for the given collection.

**The structure of an edit**
An edit is defined by specifying:

- the test to be applied,
- the domain, which is a description of the set of data that the test should be applied to, and
- the follow-up action if the test is failed.

**The test**
This is a statement of something that is expected to be true for good data. A test typically consists of data items connected by arithmetic or comparison operators. Ideas for suitable tests may arise from people with a knowledge of the subject matter, the aims of the collection or relationships that should hold between items.

Examples

- this item should not be missing
- the sum of these items equals that item

**The Domain**
The domain is defined by specifying the conditions which the data must satisfy before the test can be applied.

Example

- A test may only be relevant to those businesses in a certain industry and the domain will therefore consist of all records belonging to that industry.

**The Follow-up**
The edit designer must also think about the appropriate follow-up action if a test is failed. Some edits will be minor failures that simply require human attention, but do not need to be amended. Other edits identify major failures that require human attention and an amendment. The sort of treatment given to an edit failure is commonly done by classifying edits to a grade of severity, such as fatal, query and warning.

Example

- Where a record lacks critical information which is essential for further processing a fatal error should be displayed.

It is important to note that even if we go through comprehensive editing processes errors may still occur, as editing can identify only noticeable errors. Information wrongly given by respondents or wrongly transcribed by interviewers can only be corrected when there are clues that point to the error and provide the solution. Thus, the final computer file will not be error-free, but hopefully should be internally consistent.

Generally different levels of editing are carried out at several stages during data processing. Some of the stages involved are provided below.

### Clerical Coding

This stage includes mark-in of the forms as they are returned, all manual coding (eg. country and industry coding) and manual data conversion (eg. miles to kilometres).

### Clerical Editing

This stage includes all editing done manually by clerks before the unit data are loaded into a computer file.

### Input Editing

Input editing deals with each respondent independently and includes all "within record" edits which are applied to unit records. It is carried out before any aggregates for the production of estimates are done. An important consideration in input editing is the setting of the tolerances for responses. Setting low tolerances will result in the generation of large numbers of edit failures and impact directly on resources and in the meeting of timetables.

Ideally, an input edit system has been designed after carefully considering and setting edit tolerances, clerical scrutiny levels, resource costs (against benefits), respondent load and timing implications.

### Output Editing

Output editing includes all edits applied to the data once it has been weighted and aggregated in preparation for publication. If a unit contributes a large amount to a cell total, then the response for that unit should be checked with a follow-up.

Output editing is not restricted to examination of aggregates within the confines of a survey. A good output edit system will incorporate comparisons against other relevant statistical indicators.

### Types of Edits Commonly Used

### Validation Edit

Checks the validity or legality of basic identification or classificatory items in unit data.

Examples

- the respondent's reference number is of a legal form
- state code is within the legal range
- sex is coded as either M or F

### Missing Data Edit

Checks that data that should have been reported were in fact reported. An answer to one question may determine which other questions are to be answered and the editing system needs to ensure that the right sequence of questions has been answered.

Examples

- in an employment survey a respondent should report a value for employment
- a respondent who has replied NO to the question: Do you have any children? should not have answered any of the questions about the ages, sexes and education of any children

### Logical Edit

Ensures that two or more categorical items in a record do not have contradictory values.

Example

- a respondent claiming to be 16 years old and receiving the age pension would clearly fail an edit

### Consistency (or reconciliation) edits

Checks that precise arithmetical relationships hold between continuous numeric variables that are subject to such relationships. Consistency edits could involve the checking of totals or products.

Examples

- *totals:* a reported total should equal the sum of the reported components
  totals of different breakdowns of the same item should be equal (eg. the Australian estimate should be the same whether obtained by summing state estimates or industry estimates
- *products:* if one item is a known percentage of another then this can be checked (eg. land tax paid should equal the product of the taxable value and the land tax rate)
  income from the sales of a commodity should equal the product of the unit price and the amount sold

### Range Edit

Checks that approximate relationships hold between numeric variables that are subject to such relationships. A range edit can be thought of as a loosening of a consistency edit and it's definition will include a description of the range of acceptable values (or tolerance range).

Examples

- If a company's value for number of employees increases by more than a certain predefined amount (or proportion) then the unit will fail. Note that both the absolute change and the proportional change should be considered since a change from 500 to 580 may not be as useful as a change from 6 to 10. So, if the edit was defined to accept the record if the current value is within 20% of the previous value a change from 500 to 580 would be accepted and the change from 6 to 10 would be queried.
- In a survey which collects total amount loaned for houses and total number of housing loans from each lending institution it would probably be sensible to check that the derived item average housing loan is within an acceptable range.

This page last updated 1 March 2023

## ANALYSIS

### Introduction

In analysing results, there are some commonly used statistical methods which enable us to summarise our results. Some of these methods require a great deal of complex analysis and thus consultation with an experienced statistician is recommended, however others can be done without consulting a statistician.

The analysis that is to be carried out should be taken into account at the planning stage of the survey. Even at that early stage, the output tables should be specified and the analysis techniques decided upon to ensure that the necessary data is collected.

Once the required data from our survey has been gathered, we are in a position to carry out the analysis of our results. This involves interpreting the data, breaking it down and manipulating it to answer the survey objectives. From here the data is sorted into categories and summarised to obtain information like descriptive statistics, frequencies, percentages, correlations and measures of both location and spread. This information is used to formulate hypotheses and make inferences about our population data by estimating confidence intervals and testing the hypotheses to determine significances and trends in our data.

A useful method in analysing data involves 'getting to know it' through the use of graphical methods and summary statistics. Exploratory data analysis enforces the importance of understanding the data before determining any suitable analyses. Summarising the data using graphics and summary statistics gives a clear, immediate picture of it and reveals any key relationships between variables.

## DESCRIPTIVE STATISTICS

Descriptive statistics are summary measures which are calculated from observations in the population and are used to provide information about the distribution of particular variables in the population. In the case of responses from a sample of the population, the statistics are used to make inferences about the population.

### Condensing Data

### Forming Classes

One of the first things that needs to be done in analysing a set of observations from a population or sample is to condense the information into a more meaningful and manageable form. One fairly simple way of getting a feel for the data is to calculate a frequency distribution of responses and produce a histogram. When the data is quantitative (eg. age, income) it is necessary to group observations together into appropriate classes. The boundaries of the class intervals must be non-overlapping so that each observation can be allocated to only one interval. Obviously, the larger the class, the more information that is lost. The number of classes and the appropriate width of class interval are best determined by looking at the range of values in the data. A point worth noting is that the resulting frequency distribution is more readily comprehended if the class intervals are of equal width.

### Frequency Tables and Cross-Tabulations

A common starting point in summarising data is producing frequency counts. These simply show the number of responses for each class of the variable of interest. Frequency tables can also be used to summarise the data enabling the reader to draw his/her own conclusions. Tables can be produced for both quantitative (eg. age, income) and categorical variables (eg. sex, occupation). The data below gives an illustration of the way in which frequency counts may be used to simplify a mass of data into a univariate (one-way table):

**Employed Persons Retrenched**
**October 1990-93, Victoria**

| Age group (years) | Persons ('000) |
|---|---:|
| 18-24 | 41.7 |
| 25-29 | 40.1 |
| 30-34 | 32.1 |
| 35-39 | 35.6 |
| 40-44 | 25.3 |
| 45-49 | 22.1 |
| 50-54 | 20.7 |
| 55-59 | 16.1 |

|  |  |
|---|---:|
| 60-65 | 10.1 |
| **Total** | **244.4** |

In this case, the population covered by the data is Victorian persons aged between 18 and 65 years who were employed between October 1990 and October 1993. The variable we are looking at is 'age' and the values of ages are grouped into ranges which cover all possible ages in the population. Frequency counts for the ranges are created by counting the number of people that fall into the relevant age range. For example, there were 20,700 responses in the range 50 to 54 years old. The total of the frequency counts adds up to the total number of Victorian persons aged between 18 and 65 years who were employed between October 1990 and October 1993, which is 244,400.

It should be noted that, in a census, the frequency counts will be close to the actual number of responses in the particular ranges, with only a small adjustment needed for the non-sampling error. In a sample survey, the initial frequency counts will only be a tally of those in the sample, and will need to be inflated to provide estimates for the whole population. This process is called *weighting* the data.

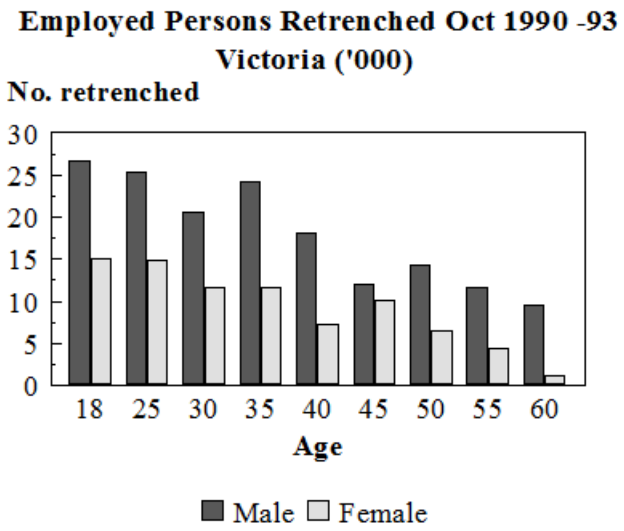
Producing a set of tables which cross-classify the key variables can give a clear picture of the data trends, and can therefore help determine the type of statistical analysis that could be undertaken on the data. A simple way to summarise the relationship between two variables is to produce a bivariate (2-way) frequency table.

**Employed Persons Retrenched**
**October 1990-93, Victoria**

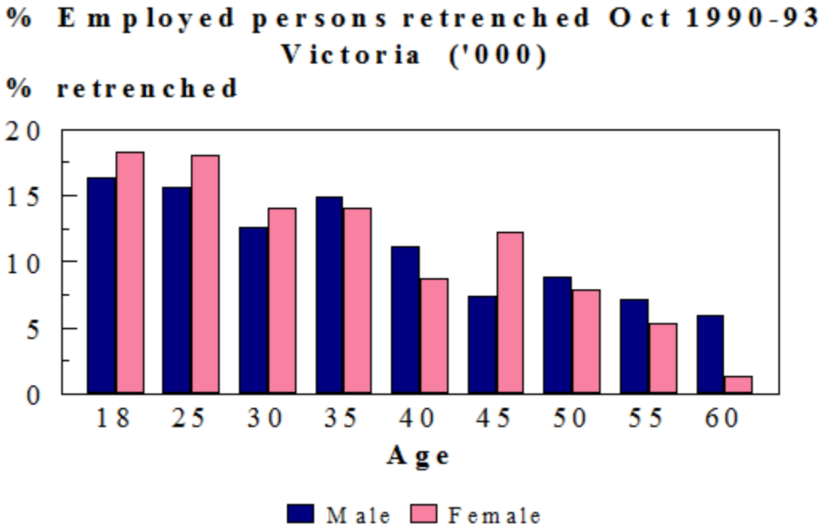
| **Age group (years)** | **Males ('000)** | **Females ('000)** | **Persons ('000)** |
|---|---:|---:|---:|
| 18-24 | 26.6 | 15.1 | 41.7 |
| 25-29 | 25.3 | 14.8 | 40.1 |
| 30-34 | 20.5 | 11.6 | 32.1 |
| 35-39 | 24.1 | 11.6 | 35.6 |
| 40-44 | 18.1 | 7.2 | 25.3 |
| 45-49 | 12.0 | 10.1 | 22.1 |
| 50-54 | 14.2 | 6.5 | 20.7 |
| 55-59 | 11.6 | 4.4 | 16.1 |
| 60-65 | 9.6 | 1.1 | 10.7 |
| **Total** | **162.0** | **82.5** | **244.4** |

For more information on tables and other forms of data presentation see Presentation of Results.

*Frequency Histograms*
Once the class frequencies have been produced, the distribution can be represented graphically by a histogram.



Sometimes instead of plotting frequencies we plot relative frequencies which show the percentage of the population within each class interval.



## Outliers

Summarising the data can be complicated if there are observations which appear to be inconsistent with the remainder of that set of data. It is important to query whether it actually belongs to the data or whether it is a computing, clerical or measurement error. An outlier is an observation that has a major effect on an estimate and which, because of its independently known atypical nature, needs special treatment. Before choosing the treatment for an outlier it is essential to know the reason for its occurrence and whether the outlier provides any information about the population of interest.

## Measures of Location (also known as Measures of Central Tendency)

It is often desirable to have summary measures to indicate the location of a frequency distribution on some sort of scale. Often the scale involved in the analysis is a time scale. This helps the researcher build up a picture of the distribution and facilitates some sort of analysis. Summary measures also enable the comparison of frequency distributions before and after a specified event (eg. number of car accidents before and after a change in traffic laws). A change can indicate a shift in the frequency distribution. The most common measures of location are the median, mean and mode.

## Mean

The mean is the most commonly used measure of location and is the average of a set of sample or population values. The formulae for calculating population and sample means are:

Population Mean

$$mean = \frac{\sum_{i=1}^{N} X_i}{N}$$

where $X_i$ is the observed value of the ith member of the population and N is the total population count.

Sample Mean

$$mean = \frac{\sum_{i=1}^{n} x_i}{n}$$

where $x_i$ is the observed value of the ith member of the sample and n is the total number of units in the sample.

For example, the mean of the set of numbers from a sample: 2, 7, 9, 11, 14 is 8.6.

The mean is used in many statistical tests (e.g. testing differences between groups) and it is possible to calculate standard errors and construct confidence intervals for the mean. In general it is the most stable measure of location but can be badly affected by extreme values.

### Median

The median is the middle value when values are sorted into order of size. If there is an even number of values in the set, the median is the average of the middle two values, for example, the median of the set of numbers: 2, 7, 9, 11, 14 is 9; while the median of the set of numbers: 2, 6, 7, 9, 11, 14 is 8. It is a good measure of location for non-symmetrical data as, in such cases, it is more central than the mean and is not affected by extreme values. It is often used in social science research, particularly in areas of housing prices and income. When analysing samples it is difficult to construct confidence intervals for the median due to the complexities in defining the sampling distribution (the distribution of the estimate of median over a number of samples).

### Percentiles

A measure of location that is linked to the median is the concept of percentiles. A percentile is a value at or below which a given percentage of the data lies. The 50th percentile is also the median as one half of the population lies below it. Two other important percentiles are the 25th percentile, known as the lower quartile boundary and the 75th percentile, known as the upper quartile boundary. The lower and upper quartile boundaries for the set of numbers: 2, 6, 7, 9, 11, 14 are 6 and 11. Percentiles can only be formed for quantitative variables.

Splitting the population into percentiles enables some comparison of the characteristics of units in each percentile. For example, the average annual income of wage and salary earners in each quartile could be compared, rather than calculating one overall average.

Percentiles are also very useful for comparing changes in characteristics of a population over time. For example, by forming income quartiles for 1986 and 1990 we can determine whether the income share of wage and salary earners in each quartile has changed over time.

### Mode

The mode of a frequency distribution is the most frequently occurring value. The mode, however, is not necessarily unique and this can cause problems in measuring the 'centre point' of our values. Having a measure of centre that is not required to take only one value can tell us more about the data than a measure like the mean or the median. In general the mean and median are better measures of location, however the mode is useful when the values are unevenly spread (eg. a two-peak distribution).

### Measures of Spread (also known as Measures of Variation)

When summarising datasets it is also important to know the variability (spread) of the values, ie. how spread out the values are around the 'centre'. A measure of location does not provide us with this information so it has to be supplemented with a measure of spread. The common measures of variability are the range, variance and standard error or standard deviation.

### Range

The range is the difference between the largest and smallest value. It is a common measure in industrial quality and meteorology because of its ease of computation. However a disadvantage of the range is that it tends to increase as sample size increases and does not provide any information about the distribution of values within the range. The range is also badly affected by extreme values. However, it can be useful for samples from small populations.

### The Interquartile Range

The interquartile range, the difference between the upper (75%) and lower (25%) quartile boundaries, covers the 'middle half' of the population. Because it does not take in the ends of the sample distribution, it is not badly affected by outliers. The interquartile range is more difficult to calculate than the range as the data has to be ranked and the quartiles calculated. The interquartile range can be used in editing to create cut-offs when deciding upon which units will be followed up to have their responses verified.

### Variance

The variance describes the spread of data around its mean and is the average of the squared differences between the value of the variable (eg. height of each person) and the mean (average height). The more the data is spread the larger the differences will be and therefore the larger the variance will be. It is possible to have two populations with the same mean but different variances. A larger variance indicates that the data is more spread out about the mean.

### Standard Deviation

The positive square root of the population variance is called the standard deviation. It is a useful measure of dispersion as it is expressed in the same units as the measures of central tendency. When applied to a sampling distribution it is referred to as the *standard error.*

### Sampling Variance

The variance of a sampling distribution is called sampling variance. The sampling variance of a statistic or estimator is simply the variance of its values obtained from all possible samples.

### Standard Error

Standard error often refers to the standard deviation of an estimator (statistic). As the value of an estimator is calculated from a particular sample, different samples will give different values of the estimator. Standard error, the positive square root of the sampling variance, measures the spread of all possible values around the mean (expected value) of an estimator. In practice, only one estimate of the "true value" is available, so the standard error can be derived mathematically. Even if the population variance is unknown, as it happens in practice, the standard error can be estimated by using the variance of the sample units. Standard error is used in the construction of confidence intervals.

### Coefficient of variation

It is the standard deviation expressed as a percentage of the sample estimate

### Correlation coefficient

The term 'correlation' refers to the existence and extent of a linear relationship between two variables. (That is, where the values of each of two variables are plotted against each other, the relationship can be approximated by a straight line.)

The correlation coefficient :

- measures the strength of the (linear) association between two variables
- ranges from +1 to -1 - (a correlation coefficient of 1 or -1 is a perfect correlation, a correlation coefficient of 0 means there is no correlation)
- does not imply causation (ie. that the values of one of the variables causes the other variable to take a specific value)
- the correlation coefficients can be tested for statistical significance under the Normal assumption

### Other Estimates

In addition to estimating means it is often of interest to measure other statistics such as totals, proportions, percentiles or even minimum and maximum values. For instance, the total turnover of retail sales for businesses in Australia or the proportion of unemployed in particular regions may be of interest.

### Further Information

The statistical analysis can comprise any summarising or presentation of the data from interpreting confidence intervals about basic summary measures calculated from the survey data, to more complex hypothesis testing using such techniques as contingency table analysis, log-linear modelling, regression analysis and time series analysis. Further information about analysis can be obtained from any of the standard texts in your library. However, it is recommended that consultation with experienced statisticians is worthwhile to determine the most appropriate analysis techniques.

## TYPES OF DISTRIBUTIONS

Arrangements of the values of a variable is called the distribution of the variable, for example, the percentages of a group of people in different age groups is called the percentage distribution of the variable 'age'. If the actual numbers or frequencies in different age groups are presented instead of percentages then it is called a frequency distribution. Similarly, the distribution which shows the probability that someone will fall into a particular age group is called its probability distribution. Therefore, a probability distribution shows the chance or probability that the value of a variable will lie in different areas within its range. The curve which shows the probability distribution is called a probability density curve.

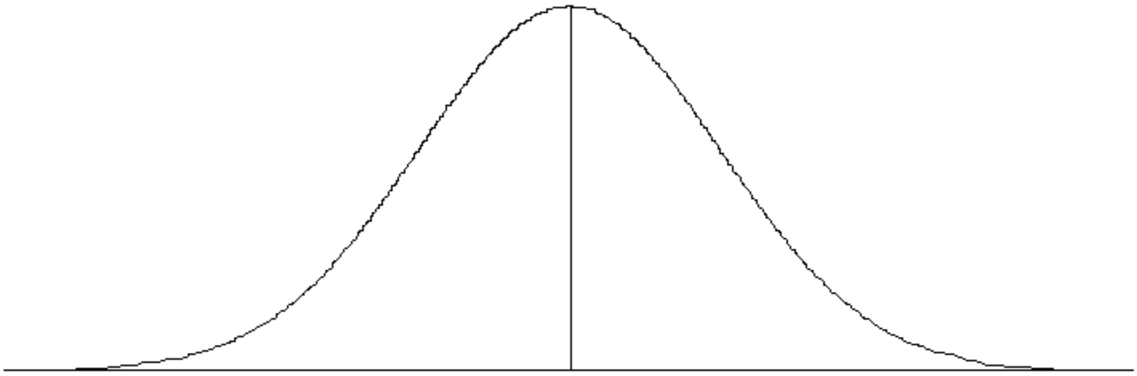
Many variables follow what is termed a Normal distribution, but other commonly sought data (eg. income) has a lopsided distribution and is said to have a skewed distribution.

### The Normal Distribution

For many populations the distributions of values is a specific bell-shaped curve, called the Normal curve. Such curves are characterised by the peak being the mean, median and mode of the distribution. The Normal distribution is the most useful distribution in statistics as most data sets can be approximated normally and nearly all sampling distributions converge to a Normal distribution. Consequently testing of hypotheses and the constructing of confidence intervals is relatively straightforward. Many phenomena in everyday life can be described by the Normal curve, for example, the height of males and females.

In a Normal distribution, 95% of all observations lie within 2 standard deviations either side of the population mean. Similarly, 99% of all observations lie within 3 standard deviations either side of the population mean. For the height example, a small proportion of a population are usually very short or very tall, with the majority falling in some middle range.

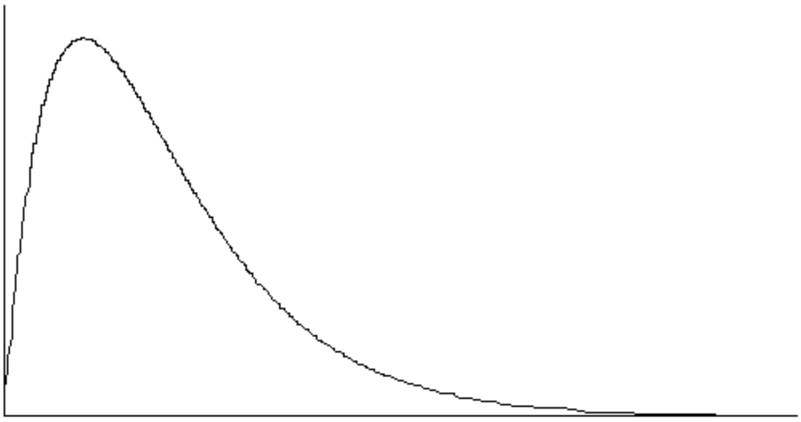
*Normal Distribution*



**Skewed Distributions**

Some populations are specified by curves that are non-symmetrical (skewed) in shape. Such curves are characterised by the peak being located to one side of the centre of the data, which causes the tails either side of the peak to be of unequal length and angle. Medians are good measures of 'location' for skewed distributions. Variables such as house prices and income are usually positively skewed.

**Positively Skewed Distribution**



## ESTIMATION

In a sample survey, we are always required to make estimates of certain population parameters. This is done in order to make inferences about the population as a whole. Good estimators are generally unbiased. In other words, theory indicates that across all possible samples the average sample value is equal to the population value, regardless of the sample size. A good estimator will also have a low variance and thus be very close to the population parameter we wish to estimate no matter which units are included in the sample that we take.

In analysing survey results, we usually require estimates of population total and mean. The two most commonly used methods of estimation are number-raised estimation and ratio estimation.

### Weighting

Note that in order to avoid the researcher drawing spurious conclusions, great care must be taken to weight and aggregate the data correctly. Weighting is the process whereby each unit in the sample has its response inflated to represent the response from all similar units in the population. The weight of a unit reflects the proportion of the population that the sampled unit represents. The weight allocated to each sample observation depends on the process used to select the sample. The most simple form of weighting is where a simple random sample (SRS) of size n is selected from a known population of size N.

**Business Example**
Suppose we want to estimate total employment for 100 cafe and restaurant businesses located in the City of Melbourne (N=100). Due to resource constraints we can only approach 10 businesses (n=10). Each business has a one in ten chance of selection and each business selected represents itself and 9 other businesses. The weight allocated to each selected business is therefore 10.

**Number-Raised Estimation**
If we observe a sequence of n observations $y_1$ , ... ,$y_n$ from a population of size N, then the number-raised estimator for the population total is the sample total multiplied by the ratio of population size to sample size (N/n).

$$y' = \left(\frac{N}{n}\right)\sum_{i=1}^{n} y_i$$

Our number-raised estimate is unbiased as the average of all possible samples is the true population total.
The basis of this method is that the average (mean) of a sample is the best estimate of the mean of a population. So if we want to find the average population of Melbourne suburbs we select a representative sample of suburbs and take the average of this sample as our estimate of the average population in Melbourne suburbs. Then to obtain the estimate of the total population in Melbourne, the average population in suburbs is multiplied by the total number of suburbs.

**Business Example cont'd**
Remember that each business has a one in ten chance of selection. If the total turnover from the sample of 10 cafe and restaurant businesses is $5 million then the number-raised estimate of total income from the population of cafe and restaurant businesses in the City of Melbourne is $5 million * 10 = $50 million.

**Advantages**
This form of estimation is easy to use and does not require any benchmark information. It is relatively simple to calculate and its variance formula is known.

**Disadvantages**
Number-raised estimation has problems in that it produces a large sampling error compared to ratio estimation and is badly affected by unrepresentative samples.

**Ratio Estimation**
Instead of calculating population values from sample values by inflating them by the ratio of the number of population units to the number of sample units, ratio estimation uses a ratio of population to sample totals based on some other variable. For example, it may be useful in a survey of job vacancies to use a ratio of total employment in the stratum to total employment in the selected firms, rather than simply the ratio of total number of firms to selected number of firms. This other variable is known as the benchmark or auxiliary variable. For it to be effective, this variable should be highly correlated with our variable of interest and needs to be known for all units in the population. If we define $y_i$ to be our variable of interest and $x_i$ as our benchmark variable, then our ratio estimate is:

$$y' = \left(\sum_{i=1}^{n} y_i\right)\frac{X}{\sum_{i=1}^{n} x_i}$$

(where X is the population total for the auxiliary variable).

The average of all possible sample estimates will not be exactly equal to the true value. Thus our ratio estimate is biased.

**Business Example cont'd**
Total employment was known to be a useful auxiliary variable in estimating total turnover of the cafe and restaurant businesses in the City of Melbourne and is known for every business in the population. The total employment for the population was found to be 1,500 people and the total employment for the sample was found to be 100 people. The calculation of the estimate of total turnover from the population uses the ratio of population to sample totals based on total employment (1,500 /100 = 15) as its weight. The ratio estimate of total turnover for cafe and restaurant businesses in the City of Melbourne is

therefore;

$5 million * 15 = $75 million.

### Advantages

The value of ratio estimation is that it decreases the standard errors of the estimates when the benchmark variable is highly correlated with the variable of interest. The ratio estimate also remains relatively unaffected by unrepresentative samples.

### Disadvantages

As we have seen, ratio estimates have the problem that they are biased. This means that, for small samples, the estimates derived may be uniformly larger (or smaller) than they should be. Ratio estimates can be less accurate than number-raised estimates if the auxiliary variable has a low correlation with the variable of interest. As a result of poor correlation, ratio estimates can also be adversely affected by outliers (unusual observations) in either the variable of interest or the benchmark variable.

### Examining Outliers

An observation should only be treated as an outlier if:

- it has a large effect on the estimates of interest and
- it is considered to be atypical and such knowledge must come from outside the dataset (for example, in a sample survey where the unit was known not to be representative of others in the same stratum).

### Example

- Estimates of Fixed Capital Expenditure may be greatly affected in a particular quarter by a single company making an atypically large purchase such as a number of passenger aircraft. In interpreting trends in the economy it would be essential to know that this was a once-off expenditure that need not reflect a general upturn in capital expenditure.

In a sample survey from a population, an outlier is rarely treated by removal. This is because every unit provides some information about the population, as the unit is itself a member of the population. Letting such a unit represent itself and no other unit is a common way of treating an outlier.

This page last updated 1 March 2023

## PRESENTATION OF RESULTS

### Introduction

Presenting the results in a clear and logical format to the client is one of the most important tasks for the person managing the survey. When presenting results, the format of the presentation should be tailored to address the aims and objectives of the survey and to satisfy the potential users of the results. Consideration should be given to the level of statistical understanding of the clients and users, particularly in regard to statistical terminology. The presentation needs to be effective, easy to understand and convey the main features of the data.

## WRITTEN REPORTS

When presenting the data, some form of a written report is essential. The report should convey the main features clearly and follow a logical progression, use as little jargon as possible, provide insight into the data and make the results as interesting as possible. As the author of the report, you should convey specific messages rather than generalised information. You should also put forward theories to explain the findings and encourage the users to further explore the data. If appropriate, you may need to include recommendations.

The contents of the report and its balance of words, tables and graphs will naturally depend on the topic, the results and the likely readers. Graphs and tables convey complex information clearly and can be used to add variety. At all times it should be remembered that the report is written to be read, and so needs to make sense and be understandable.
A survey report generally covers:

### Introduction

The introduction states the purpose and aims of the survey and the aims of the report; gives the background to the research; defines terms and concepts; and states whether the survey is testing an hypothesis or is exploratory.

### Methodology

The methodology describes the method of sampling and information on the survey population, as well as how the data was analysed and the statistical procedures which were used.

### Findings and Analysis

The findings and analysis section is the main part of the report which deals with details of the sample numbers, response rates, results and interpretation of tabulations etc. and discusses possible courses of action.

### Conclusions and Recommendations

The conclusions summarise the major findings of the report and answers questions posed in the introduction. The recommendations outline what actions are indicated on the basis of the conclusions.

### Appendices and References

The appendices consists of items which may be useful to the reader (e.g. the questionnaire) but not essential to the report. References list the books, journals and papers referred to in the study.

Remembering that the report is likely to serve as a basis for discussion, some other important considerations are the title, use of headings and sub-headings, the colour and design of the cover and the overall appearance of the report - it should stimulate the reader's interest.

## STATISTICAL PRESENTATION

The manner in which the results are presented will depend on the data and the types of descriptive statistics required. Tables and graphs are the most common form of presentation but other types are available. In general, tables are more detailed - showing the actual values, whereas graphs are more useful in showing relationships - concentrating on the form, shape and movement of the data. Graphs are particularly useful in representing the change in the value of a data item over a period of time. Usually either a graph or a table will be sufficient to represent a set of data.

### Tables

Tables are the most common form of statistical presentation. Tables present additional information which cannot be shown in general text. A good table is one in which patterns and exceptions stand out when looking at the table, followed by a small paragraph commenting on the table. Tables are usually only used to present a few values as they become difficult to comprehend if they involve too many numbers.

Tables should have clear headings. Some other useful guidelines are:

- to round data to 4 or 5 significant figures which makes the data easier to see and manipulate;
- reading down is easier than reading across a row, especially for a large number of items;
- row and column averages or percentages may help the reader interpret the data - widely spaced columns are difficult to compare and should be avoided; and
- totalling of rows and/or columns is usually helpful.

## Graphs

We can also use a sequence of simple graphs to tell a story about the data and to give insight into findings. This can be supported by an explanatory paragraph.

Graphs are a visually attractive way of presenting data. They maintain the interest of the readers and encourage them to think about the data. Although the amount of information which can be presented is limited, graphs often reveal or highlight "hidden" facts in complex data. Trends and relationships are more clearly grasped from a graph rather than a table or text and are therefore better remembered by the reader. However, graphs are not always necessary and should only be used where it is appropriate. Keep in mind that the reader may not be familiar with the data and therefore you need to include sufficient information on the graph to explain the data.

## Constructing Graphs

- Titles: A title is essential and is best placed at the top. It should indicate the "what, where and when" of the graph as concisely as possible and should be larger than the lettering on the graph itself.
- Scales: the horizontal scale usually measures the time unit where a graph is over a period of time, and the vertical scale measures the variable under consideration. Where possible the vertical scale should begin at zero.
- shading and colouring can be added to create visual emphasis;
- both axes should be marked and named with scales and units;
- all parameters and variables should be defined as concisely as possible;
- a legend may be necessary, particularly for detailed graphs with shading and/or colouring;
- labels should be used if more than one variable is plotted on the same graph; and
- footnotes may be used to explain unusual features such as breaks in the series.

## Types of Graphs

The decision as to which type of graph to use depends on the type of data being presented. The following are some of the types of graphs that could be used:

- Line or Curve Graphs
  These graphs are used when emphasis on the movement rather than the size of the data item or when several series are being compared. These types of graphs show variations in the data plotted over a period of time.
- Bar or Column Graph
  This type of graph depicts numerical values over a given variable. The value is represented by the height of the column. This type of graph is especially effective for showing large changes from one period to the next.
- Grouped Columns/Bars
  These graphs compare different categories on the same graph. Groups of columns, with each column in the group representing a different category, are plotted on the same axis.
- Pie Chart
  These charts are used for comparison (in percentage terms) of components. The components can be compared with each other and the contribution they make with the whole. These graphs are best used when there is a small number of categories.
- Map
  A map is used to show boundaries of areas. Various shadings for different regions can show how the value of a variable changes between locations.
- Other Graphs
  Some other graphs which are available include: Index, Pictogram, Surface Chart, Cumulative Curve, Deviation Line, Sliding Bar Chart, Dot Chart, Population Pyramid, Map Chart, Stacked Bar and Scatter Diagram.

## Other Forms of Presentation

### Oral Presentations

Depending on the circumstances (users, type of data, results), a written report may be inadequate or may need to be supplemented. Oral presentation of the results of a survey is often neglected as an important means of conveying information. Whereas a written report provides great detail with a wide range of results, an oral session can only emphasise a few major points. However, this can often be most suitable depending on the audience. As with a written report, poor presentation may cause the survey results to be rejected. The spoken word and visual aids can have a great impact on an audience. The presenter therefore should be aware of who the presentation is aimed at and know what survey results may be contrary to existing ideas.

### Posters

A poster is one way to attract attention but only one direct statement should be made. The message must be noticeable at a glance and the poster itself must be attractive to encourage possible users to inspect it.

### Panel Exhibits

A panel exhibit is an extension of the poster presentation. This type of presentation gives more details and expands several main ideas. Again it is important that the panels be colourful and attractive.

### Charts

A chart maybe appropriate depending on the data. A chart is an outline map which can contain photographs, words, figures, graphs, diagrams, maps, symbols, illustrations, etc (eg. a weather chart).

### Videos

The use of videos and television can provide an additional means of communication of the survey results.

## Conclusion

It is important to present data well. If we have gone to the effort, time and expense of undertaking a study, then we have to ensure that we present the data in a clear, logical way. We must consider the use of text, tables and diagrams to construct a document that is useful and understandable. The ultimate aim of any presentation is to inform the user of the results of the research in a way that can be used to promote informed decision making.

This page last updated 1 March 2023

## CONFIDENTIALITY

### Introduction

Confidentiality is the act of protecting the privacy of respondents through ensuring that information collected is not revealed to a third party. This includes ensuring that a third party does not sight the form filled out by a respondent, and also ensuring that the publishing or disclosing of data cannot be related back to individual respondents. Assurances of confidentiality can be provided in an advance letter to the respondent, on the form itself or by the interviewer. If respondents are informed that a collection agency is taking steps to ensure confidentiality, response rates can be improved.

### Factors to Consider

Agencies involved in the collection of data have a responsibility to ensure the protection of the information that they have been supplied with. Collectors should implement procedures to ensure confidentiality. There are a number of ways in which the privacy of respondents can be maintained:

- Code numbers, instead of names, can be used on the questionnaires to minimise the links that can be made between questionnaires and respondents.
- Suppression of personal identification information such as names, addresses and telephone numbers should be undertaken at the data processing or analysis stage.
- Questionnaires and data sources can also be destroyed as early as possible in the process.
- When creating tables, broad cross-classifications can be used to avoid cells with only a small number of contributing units.

### Securing of Forms

Forms should be locked up in a safe place and destroyed not long after the project is finished. If unit record tapes (magnetic tapes containing the complete survey records for each individual respondent) are to be released then the following things must happen: a third party must not be able to recognise an individual or an organisation from that release; and that third party must not learn something that they did not already know about the individual or organisation.

There is an argument for keeping forms for a long period of time and then releasing the complete information kept in them when the respondents are no longer alive. Consideration in this case must be given to descendants of respondents.

### Publishing Results

When presenting or publishing data in the form of tables, confidentiality considerations need to be taken into account especially when there exists cells with small counts (for example, cells representing only two observations). There are many techniques which can be used for confidentialising tabulations; some of which are described below.

## TYPES OF DISCLOSURE

The aim of confidentialising data prior to publication is to prevent the disclosure of information where that information would identify the respondents who provided the data. There are several types of disclosure that can occur.

### Direct Disclosure

Direct disclosure occurs as a direct result of a given cell value in a statistical table. Cells leading to direct disclosure are called sensitive cells. To determine which cells are sensitive, certain rules are generally used - the threshold rule for count data and the (n,k) dominance rule for quantities.

### *I*nadvertent Disclosure

Although a cell value in a statistical table may be suppressed, the value can be worked out through the addition or subtraction of other cell values and the totals. Consideration of this kind of disclosure leads to the problem of consequential (or complementary) confidentiality.

### Residual Disclosure

Residual disclosure results from a comparison of a set of tables. In this situation, the differences between the tables may provide details on a particular information source. Consideration of this type of disclosure leads to complex confidentiality problems.

### Disclosure from External Information

This type of disclosure occurs from a comparison with information obtained from a source which is independent of that from which the tables are obtained (that is, an external source). There is no systematic method available to protect against such disclosure.

## TYPES OF CONFIDENTIALITY ASSURANCE TECHNIQUES

There are several techniques that have been developed to minimise the risk of disclosure of information that can be traced back to the responding units. These techniques fall into three main categories, as listed below.

### Data Suppression

This technique simply involves not releasing that information which may identify individuals. This is probably the oldest method but it still offers interesting avenues of investigation in terms of automation. To date some simple automated suppression algorithms have been developed for two dimensional tables. The far more complex problems associated with tabulations of higher dimensions are still being tackled. To produce an efficient and practical automated system requires high resource input and much fine tuning.

### Threshold Rules and Cell Concentration Rules

A threshold rule specifies the minimum number of units that must contribute to the value of a cell. Where the number of units contributing to the value of a cell is less than a pre-specified threshold value, then the cell would be suppressed in order to prevent disclosure.

The cell concentration rule (also called a cell dominance rule) prevents the publication of cells where a small number of respondents contribute a large percentage to the cell total. For example, it may be decided that if two respondents contribute 85 per cent to a cell total, the cell will not be published.

### Data Rounding

Random rounding involves the technique of replacing small values that are to appear in a table with other small random numbers. Because random rounding results in data distortion, it is therefore not additive (additivity means that the table total, either between or within tables, are equal to the sum of the relevant cell values or subtotals). This technique can be unbiased (a value is biased if the expected value of the data after a confidentiality techniques has been applied does not equal the value of the original entry it is replacing) if done in an appropriate manner.

Controlled rounding is a combination of conventional rounding and random rounding. Controlled rounding may result in additivity, unbiasedness, and reduction in data distortion (when compared to other rounding methods). However, this method may not give consistency among tables.

### Conclusion

Since any release of statistical information will result in something being learnt about all units in that release, this policy could be interpreted literally as meaning that no statistics should be published. Obviously, a reasonable confidentiality policy involves establishing an acceptable level of risk of identifying individuals and balancing this against the amount of legitimate statistical information contained in a release.

This page last updated 1 March 2023